Project 2475 | April 2026

SJSU SAN JOSÉ STATE UNIVERSITY

MINETA TRANSPORTATION INSTITUTE

Spatiotemporal Dynamics of Daily and Per Capita VMT in California: A County-Level Analysis (2019–2023)

Yong Lao, PhD
Bo Yang, PhD

California State University Transportation Consortium

California State University MONTEREY BAY

MINETA TRANSPORTATION INSTITUTE

Founded in 1991, the Mineta Transportation Institute (MTI), an organized research and training unit in partnership with the Lucas College and Graduate School of Business at San José State University (SJSU), increases mobility for all by improving the safety, efficiency, accessibility, and convenience of our nation's transportation system. Through research, education, workforce development, and technology transfer, we help create a connected world. MTI leads the California State University Transportation Consortium (CSUTC) funded by the State of California through Senate Bill 1 and the Climate Change and Extreme Events Training and Research (CCEETR) Program funded by the Federal Railroad Administration. MTI focuses on three primary responsibilities:

Research

MTI conducts multi-disciplinary research focused on surface transportation that contributes to effective decision making. Research areas include: active transportation; planning and policy; security and counterterrorism; sustainable transportation and land use; transit and passenger rail; transportation engineering; transportation finance; transportation technology; and workforce and labor. MTI research publications undergo expert peer review to ensure the quality of the research.

Education and Workforce Development

To ensure the efficient movement of people and goods, we must prepare the next generation of skilled transportation professionals who can lead a thriving, forward-thinking transportation industry for a more connected world. To help achieve this, MTI sponsors a suite of workforce development and education opportunities. The Institute supports educational programs offered by the Lucas Graduate School of Business: a Master of Science in Transportation Management, plus graduate certificates that include High-Speed and Intercity Rail Management and Transportation Security Management. These flexible programs offer live online classes so that working transportation professionals can pursue an advanced degree regardless of their location.

Information and Technology Transfer

MTI utilizes a diverse array of dissemination methods and media to ensure research results reach those responsible for managing change. These methods include publication, seminars, workshops, websites, social media, webinars, and other technology transfer mechanisms. Additionally, MTI promotes the availability of completed research to professional organizations and works to integrate the research findings into the graduate education program. MTI's extensive collection of transportation-related publications is integrated into San José State University's world-class Martin Luther King, Jr. Library.

Report 26-07

Spatiotemporal Dynamics of Daily and Per Capita VMT in California: A County-Level Analysis (2019–2023)

Yong Lao, PhD

Bo Yang, PhD

April 2026

A publication of the
Mineta Transportation Institute
Created by Congress in 1991

College of Business
San José State University
San José, CA 95192-0219

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. 26-07	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle Spatial-Temporal Dynamics of Daily and Per Capita VMT in California: A County-Level Analysis (2019–2023)		**5. Report Date** April 2026
		6. Performing Organization Code
7. Authors Yong Lao, PhD Bo Yang, PhD		**8. Performing Organization Report** CA-MTI-2475
9. Performing Organization Name and Address Mineta Transportation Institute College of Business San José State University San José, CA 95192-0219		**10. Work Unit No.**
		11. Contract or Grant No. SB1-SJAUX_2023-26
12. Sponsoring Agency Name and Address State of California SB1 2017/2018 Trustees of the California State University Sponsored Programs Administration 401 Golden Shore, 5th Floor Long Beach, CA 90802		**13. Type of Report and Period Covered**
		14. Sponsoring Agency Code
15. Supplemental Notes 10.31979/mti.2026.2475		

16. Abstract

Vehicle miles traveled (VMT) is a fundamental metric for assessing mobility trends and infrastructure needs. This study examines the spatial-temporal dynamics of daily VMT (DVMT) and per capita DVMT across California counties from 2019 to 2023, covering the pre-, mid-, and post-pandemic periods via GIS mapping and k-means clustering. To identify determinants of per capita DVMT, we compared traditional linear regression approaches (OLS, Ridge, LASSO, Elastic Net) with ensemble tree-based models. Specific results include: the ensemble models estimated using 2019–2022 data delivered substantially higher accuracy, achieving R^2 values exceeding 0.98; meanwhile, out-of-sample performance on 2023 data remained robust ($R^2 \approx 0.82$ for Random Forest; $R^2 \approx 0.91$ for Gradient Boosting), indicating strong model generalizability. Feature importance analysis identifies housing density, population density, and public transit mode share as the primary drivers of per capita DVMT. These findings underscore the utility of spatial analysis and advanced nonlinear modeling for regional transportation planning.

17. Key Words Vehicle miles traveled, k-means clustering, regression analysis, transportation GIS	**18. Distribution Statement** No restrictions. This document is available to the public through The National Technical Information Service, Springfield, VA 22161.		
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 68	**22. Price**

Form DOT F 1700.7 (8-72)

DOI: 10.31979/mti.2026.2475

Mineta Transportation Institute
College of Business
San José State University
San José, CA 95192-0219

Tel: (408) 924-7560
Fax: (408) 924-7565
Email: mineta-institute@sjsu.edu

transweb.sjsu.edu/research/2475

Acknowledgments

The authors are grateful to the Mineta Transportation Institute for funding this research. We would like to thank Lisa Rose and Editing Press for editorial services, as well as MTI staff Rajeshwari Rajesh and Katerina Earnest for project management and graphic design assistance.

CONTENTS

LIST OF FIGURES

List of Tables

Executive Summary

Vehicle miles traveled (VMT) is a central indicator for transportation system performance, climate impacts, and infrastructure planning in California. This study examines county-level daily vehicle miles traveled (DVMT) and per capita DVMT from 2019 to 2023 to assess how travel behavior changed before, during, and after the COVID-19 pandemic. Using GIS-based cluster analysis alongside linear and machine learning models, the research provides a comprehensive assessment of spatial-temporal VMT patterns and their key drivers.

The results show that statewide DVMT and per capita DVMT declined sharply in 2020 (-14.9%) and remained below pre-pandemic levels by 2023, indicating partial structural changes in travel behavior rather than a full rebound. Spatial analysis reveals strong regional heterogeneity: large metropolitan and tech-oriented counties experienced the largest and most persistent declines, while inland logistics hubs and rural recreational counties showed greater resilience in DVMT. Time-series clustering identified distinct county-level travel profiles, underscoring the limitations of uniform statewide VMT policies.

To explain cross-county variation in per capita DVMT, the study compares four linear regression models (OLS, Ridge, LASSO, Elastic Net) with two tree-based ensemble methods (Random Forest and Gradient Boosting). While linear models explain approximately 47% to 50% of the out-of-sample variance, ensemble models achieve substantially higher predictive accuracy (test $R^2 \approx 0.82–0.91$), indicating the presence of nonlinear relationships and interaction effects. Across all models, housing intensity (housing units per 1000 people) and population density are the dominant determinants of per capita DVMT, whereas public transit mode share plays a consistent but secondary role. Income, poverty, and commute time explain less variance once built environment factors are controlled for.

Overall, these findings underscore the value of integrating GIS-based spatial analysis with advanced machine learning to inform transportation and climate policy. The results suggest that effective VMT reduction strategies should prioritize land use patterns, housing development, and multimodal accessibility, while accounting for regional variation across California. The study provides evidence to support more spatially targeted, equitable, and data-driven approaches to transportation planning under frameworks such as SB 375 and SB 743.

1. Introduction

Vehicle miles traveled (VMT) is estimated by counting vehicles passing specific locations and multiplying those counts by roadway segment length to determine the total miles driven. It has emerged as the fundamental metric for assessing transportation system performance and environmental impacts in California. In 2006, California passed the landmark Global Warming Solutions Act (AB 32), setting an ambitious target to reduce greenhouse gas (GHG) emissions to 80% below 1990 levels by 2050. Further, Senate Bill 375 (2008) advanced this agenda by explicitly targeting VMT. Subsequently, Senate Bill 743 (2013) mandated replacing the traditional Level of Service (LOS) metric with VMT as the primary measure for evaluating transportation impacts under the California Environmental Quality Act (CEQA). To comply with SB 743, the Governor's Office of Planning and Research (OPR) selected a VMT threshold that supports state climate goals, with technical advisory documents suggesting that achieving a 15% reduction in VMT at the project level is necessary and feasible across various place types (Governor's Office of Planning and Research, 2018). Despite these legislative efforts, before the COVID-19 pandemic, both total and per capita VMT in California were generally increasing or plateauing rather than decreasing consistently toward the ambitious 15% reduction goal. The state temporarily met VMT-reduction targets in 2020 due to the COVID-19 pandemic. Since then, California's annual VMT reached a record high of 344.0 billion miles in 2023 (Caltrans, n.d.). Additionally, per capita VMT has fluctuated rather than declined consistently.

Previous research has identified a wide range of factors influencing VMT, such as population density, land use, household income, job-housing balance, transit accessibility, and policy interventions (Holtzclaw et al., 2002; Salon et al., 2012; McMullen & Eckstein, 2013; Woldeamanuel & Kent, 2014; Circella et al., 2016; Dong, 2022). Despite this foundational work, a critical gap remains: the existing literature has not fully examined the complex geographical variation in VMT and the spatially non-uniform nature of its interaction with different socioeconomic factors. Substantial heterogeneity exists in VMT patterns across California, driven by wide variation in land use, economic structure, transportation infrastructure, and demographic composition. Urban coastal counties, inland metropolitan regions, and rural areas often exhibit markedly different VMT levels, growth trajectories, and responses to socioeconomic change. In a large, demographically and geographically diverse state like California, failing to account for these regional patterns severely limits the effectiveness of statewide transportation and climate policy. Policies designed around statewide averages risk overlooking local dynamics, leading to misaligned investments and uneven policy outcomes.

This study contributes to the literature by bridging the gap between statewide policy goals and regional spatial realities. By utilizing Geographic Information Systems (GIS) and advanced spatial modeling, we move beyond aggregate averages to reveal the "where" and "why" of VMT fluctuations. The significance of this research lies in its ability to identify distinct mobility archetypes and quantify location-specific drivers of travel behavior. Such insights are vital for developing targeted, efficient, and equitable transportation strategies that acknowledge the unique constraints of different California jurisdictions. To this end, we address three primary research questions:

How did DVMT and per capita DVMT change across California counties between 2019 and 2023, and how did those patterns differ across regions?

Are there distinct spatial-temporal mobility archetypes based on counties' pre-, mid-, and post-pandemic trajectories?

Which socioeconomic and built environment factors most strongly predict cross-county variation in per capita DVMT, and do nonlinear machine learning models outperform traditional linear regression models?

We employ GIS to visualize these intricate VMT patterns and utilize a combination of statistical and machine learning techniques to rigorously assess the relationships between VMT and a comprehensive set of socioeconomic factors. By providing a spatially explicit framework, this study offers a more granular and precise evidence base for planners seeking to meet the state's ambitious climate targets.

The remainder of this report is organized as follows. Section 2 reviews the literature on VMT factors and analytical models. Section 3 discusses the data sources and methods used for GIS-based mapping and cluster analysis. Section 4 presents our linear regression and machine learning models along with detailed results and interpretations. Finally, Section 5 offers the conclusions and future research directions.

2. Literature Review

Vehicle miles traveled (VMT) is a key indicator of transportation demand, environmental externalities, and infrastructure use. Explaining its spatial variation and temporal dynamics requires integrating evidence across socioeconomic, spatial, and behavioral domains. This literature review synthesizes peer-reviewed research, federal and state technical reports, and recent empirical studies to examine two overarching themes: (1) the structural determinants of VMT across geographic contexts, and (2) the quantitative methods used to estimate, analyze, and forecast VMT at multiple scales. The following sections review the empirical evidence within each theme.

2.1 VMT-Related Socioeconomic and Spatial Factors

Vehicle miles traveled (VMT) varies geographically and is influenced by a wide range of factors. First, socioeconomic factors such as GDP, employment, income, housing, and demographics set the broad trajectory of VMT, but urban planning and policy choices can significantly modulate these forces. Historically, VMT has been used to track economic activity. A Federal Highway Administration (FHWA) report suggested a long-run association between GDP and total VMT, with periods of "decoupling" (i.e., slower VMT growth relative to GDP) emerging in the 2000s (FHWA, 2012). A study by McMullen and Eckstein (2013) examined factors affecting vehicle miles traveled (VMT) in 87 U.S. urban areas between 1982 and 2009. It found that higher lane-miles, per capita income, and employment in the construction and public sectors were associated with increased VMT. Conversely, factors such as fuel costs, transit use, and employment in manufacturing, retail, and wholesale sectors were linked to decreased VMT. Income is the primary determinant of vehicle ownership, and increases in income tend to raise household VMT via higher car ownership and more discretionary travel (Goodwin et al., 2004; Salon, 2016; Mondschein & Taylor, 2017). However, low-income households may convert rising income into greater mobility at a faster rate than higher-income households (Blumenberg & Pierce, 2012). Housing markets shape where households live relative to jobs and amenities, thereby influencing commute distances and VMT (Blumenberg & Siddiq, 2023). Program evaluations and policy briefs have linked higher employment density and transit-oriented housing to lower per capita VMT (Circella et al., 2020; Urban Displacement Project, 2021).

Second, the built environment plays a crucial role: urban density and compact development are associated with lower VMT (Zhang et al., 2012; Ewing et al., 2013; Ewing et al., 2016), while suburban and rural areas tend to have higher VMT due to factors such as longer travel distances (Zhou et al., 2021). For car travel demand, the estimated long-run elasticity with respect to population density is -0.8 (Litman, 2025). Other characteristics, such as land use mix and accessibility, also significantly explain variation in VMT. By locating employment centers, retail establishments, and housing in proximity, mixed-use neighborhoods shorten vehicle trips and encourage non-motorized travel modes, thereby reducing passenger vehicle GHG emissions

(Ewing & Cervero, 2010; Diao & Ferreira, 2014; Hong et al., 2014; Boarnet & Wang, 2019; Su et al., 2022).

Third, transportation infrastructure and public transit are both critical components. Increased freeway capacity is associated with higher VMT, while better transit service is linked to lower VMT (Ewing et al., 2013; Chi & Lee, 2024; Sabouri et al., 2024; Millard-Ball & Rosen, 2025). Conversely, traffic congestion has been found to reduce VMT (Sardari et al., 2018). Transit access, in particular, shows significant reduction benefits. One analysis of four major California metropolitan areas found that households within 0.5 miles of rail transit stations had lower daily VMT and greenhouse gas emissions across virtually all income levels (Boarnet et al., 2018). Another study attributed approximately 15% of the overall VMT reduction to transit access, although it noted that the magnitude of this effect varied substantially with the age and connectivity of rail systems (Alexander et al., 2021). It is estimated that households in transit-oriented development (TOD) zones in California, on average, save 18% of their total annual transportation expenditures (Dong, 2022).

Fourth, COVID-era disruptions and durable increases in remote and hybrid work altered the income–employment–VMT linkage (Reiffer et al., 2023). Fisher and LaMondia (2021) analyzed how temporal, regional, demographic, and policy factors influenced VMT across the contiguous United States during the early stage of the COVID-19 pandemic. They found that urban and densely populated counties experienced substantially larger reductions in traffic volume than rural areas. Zheng et al. (2024) quantified near-one-for-one VMT reductions with fewer on-site workers at the state level during 2020–2022, with heterogeneous effects across states. Meanwhile, FHWA projections continue to condition medium-term VMT growth on macroeconomic scenarios, demographics, and behavioral trends (Pickrell et al., 2023).

Ultimately, socioeconomic and spatial factors interact in complex, scale-dependent ways to influence VMT. A four-way interaction was detected among macroscale and microscale built environment features, transit access, and income, indicating that VMT depends on complex combinations of factors rather than on individual elements (Alexander et al., 2021).

2.2 Quantitative Techniques to Estimate, Analyze, and Forecast VMT

Given that VMT is the key metric for understanding traffic congestion, air pollution, and gas tax revenue, and because it plays a vital role in how policymakers allocate transportation funds, numerous studies have been conducted on estimating, analyzing, and forecasting VMT at various geographical scales (Liu et al., 2006; Williams et al., 2016; Klakto et al., 2017; Pickrell et al., 2023). These studies employ diverse mathematical and statistical modeling techniques, with geographic scales ranging from fine-grained 250×250m grid cells and individual sites to metropolitan areas, state-level analyses, and national coverage.

Multiple linear regression was commonly used to model the relationship between VMT and a variety of independent variables, such as socioeconomic factors (McMullen & Eckstein, 2013), the sprawl of metropolitan statistical areas (MSAs) (Gerayeli & Jenkins, 2016), the built environment (Alexander et al., 2021), and CO_2 and NO_2 emissions (Chung et al., 2024). Moreover, spatial considerations necessitated more sophisticated approaches in several studies. Diao and Ferreira (2014) employed spatial regression models, specifically comparing spatial-lag and spatial-error models with the ordinary least squares (OLS) models. The spatial-error model proved to be a better fit, addressing heteroskedasticity and spatial autocorrelation problems. Kim et al. (2016) applied spatial interpolation techniques (i.e., regression kriging and linear regression) in predicting the annual average daily VMT for unmeasured locations in the urban area of Bucheon, South Korea. Bamney et al. (2021) employed a two-way random effects linear regression model to assess changes in county-level VMT during the COVID-19 pandemic. Schaber et al. (2024) used mobile phone data from 26 U.S. metropolitan areas to examine whether a ZIP code's geographic location or its sociodemographic characteristics better predicted changes in daily mobility during the COVID-19 pandemic. The researchers found that while specific zip-code characteristics (like median income and population density) helped explain variation within a city, the city itself was the primary driver of deviations from national mobility trends. This is likely because the city variable inherently captures the unique sociodemographic composition of that area.

Structural equation modeling (SEM) has been used in several studies to capture complex relationships among factors. Ewing et al. (2013) applied SEM to explain VMT levels in 315 urbanized areas, while a subsequent study (Ewing et al., 2014) used SEM to model both cross-sectional and longitudinal changes in VMT. Nasri and Zhang (2014) employed SEM to analyze data from 19 U.S. metropolitan areas, examining how the built environment at local, regional, and metropolitan scales causally influences travel behavior while controlling for residential self-selection. The study found that regional-level factors, such as the presence of employment subcenters and overall regional accessibility, have a more significant impact on reducing driving than local neighborhood policies alone. These SEM approaches enabled researchers to quantify both direct and indirect effects, such as the land use multiplier effect of transit on VMT.

Machine learning methods have emerged as powerful tools for predicting VMT in recent years. Zhou et al. (2021) compared three machine learning approaches: Gradient Boosting Machine (GBM), Linear Regression, and Random Forests. The research used census tract-level data, combining household income with annual VMT, vehicle fuel economy, and fuel costs to provide a granular understanding of energy affordability across the United States. The study found that the household transportation energy burden varies significantly across regions and socioeconomic groups, with lower-income households and suburban/rural areas facing higher burdens. The GBM model outperformed the other methods across 18 regions. In a study by Tian et al. (2024), a comparative analysis was conducted between traditional multilevel regression models and boosted regression trees (BRT). The researchers used household travel survey data, including sociodemographic variables (i.e., age, employment status, household size, income), regional climate data, metropolitan population, and average gas prices. The study found that while both approaches were effective in demonstrating the impact of land use variables, the machine-learning model offered superior predictive accuracy and the ability to identify nonlinear impacts and useful planning thresholds. Kalantari et al. (2025) applied a one-vs.-rest Random Forest machine learning model to analyze over 800,000 trips across 29 U.S. regions, comparing its predictive performance with traditional nested and multinomial logit models commonly used by metropolitan planning organizations. The RF model significantly outperformed conventional methods in accuracy, identifying travel time, distance, and vehicle ownership as the primary determinants of mode choice. It also demonstrated that built environment features, such as activity and transit density, are crucial for encouraging sustainable transportation.

From the literature review, it is clear that understanding the spatial patterns and dynamics of vehicle travel requires a robust analysis of defining data sources, key metrics, and overarching trends. Consequently, this study focuses on county-level daily vehicle miles traveled (DVMT) from 2019 to 2023 to characterize the state's transportation landscape. The methodology employs two primary approaches: (1) GIS-based mapping and cluster analysis; and (2) statistical and machine learning modeling.

3. Spatial Analysis and Mapping of DVMT

3.1 Data and Study Scope

DVMT represents the total number of miles traveled by all motor vehicles within each county on an average day. The primary source of this data is the Highway Performance Monitoring System (HPMS), a comprehensive, federally mandated program in California (Caltrans, n.d.). Baek and Kim (2023) developed a detailed Transportation Data Guide for local governments in California to access and utilize local VMT and related data.

County-level explanatory variables included population, gross domestic product (GDP), income, vehicle ownership, housing units, and commuting patterns (Table 1). These predictors were selected to capture demographic, economic, and behavioral dimensions known to influence vehicle travel demand. Real GDP data were obtained from the Bureau of Economic Analysis (BEA) website. All the other data were drawn from the Census Bureau's 2019–2023 American Community Survey (ACS) 5-Year Estimates (released in 2024).

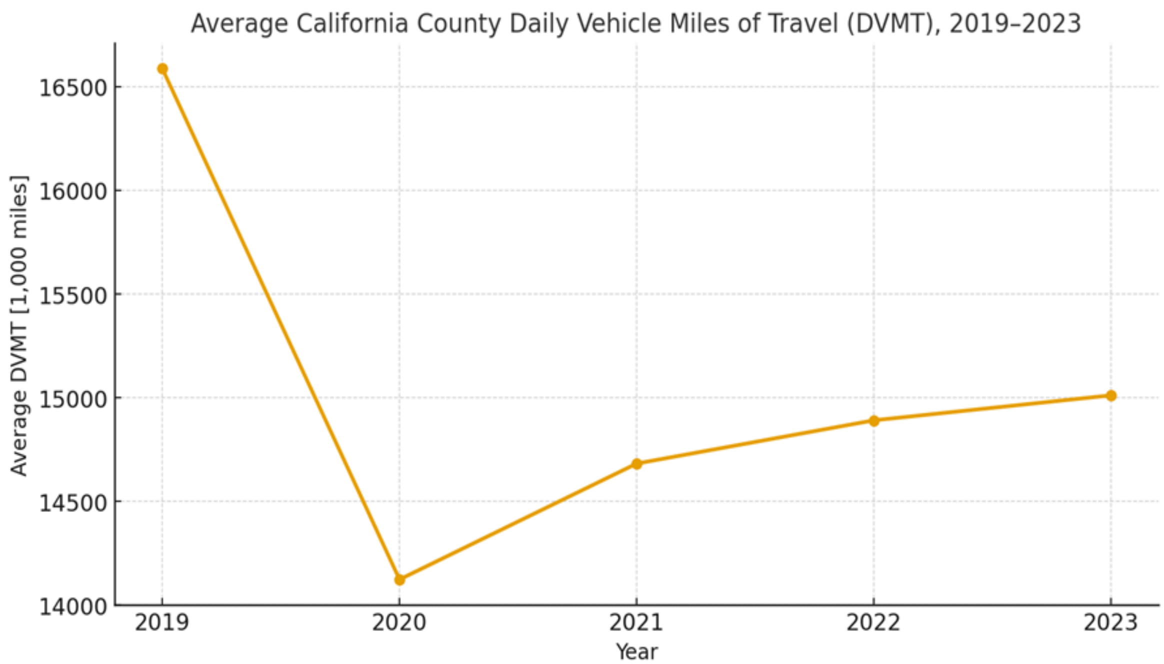
Table 1. County-Level Variable Descriptions

Variables	**Descriptions**
Daily Vehicle Miles Traveled	Average daily VMT (in 1000 miles)
Population	Total resident population of the county
Population density	Number of residents per square mile
Real GDP	Inflation-adjusted value of goods and services produced within the county
Civilian employed population 16 years and over	Estimated number of employed civilians in the county
Median household income	Median household income (in dollars)
Population with income in the past 12 months below the poverty level	Total number of individuals living below the poverty level
Percentage of the population below the poverty level	Percentage of individuals living below the poverty level
Housing units	Total number of housing units in the county
Housing units per 1000 people	Number of housing units per 1000 residents
Aggregate number of vehicles (car, truck, or van) used in commuting	Total number of vehicles used for commuting
Aggregate travel time to work	Total job-related commuting time (in minutes)
Percentage of workers 16 years and over commuting by public transportation	Percentage of workers utilizing public transit for commuting

3.2 Spatial-Temporal Dynamics of DVMT

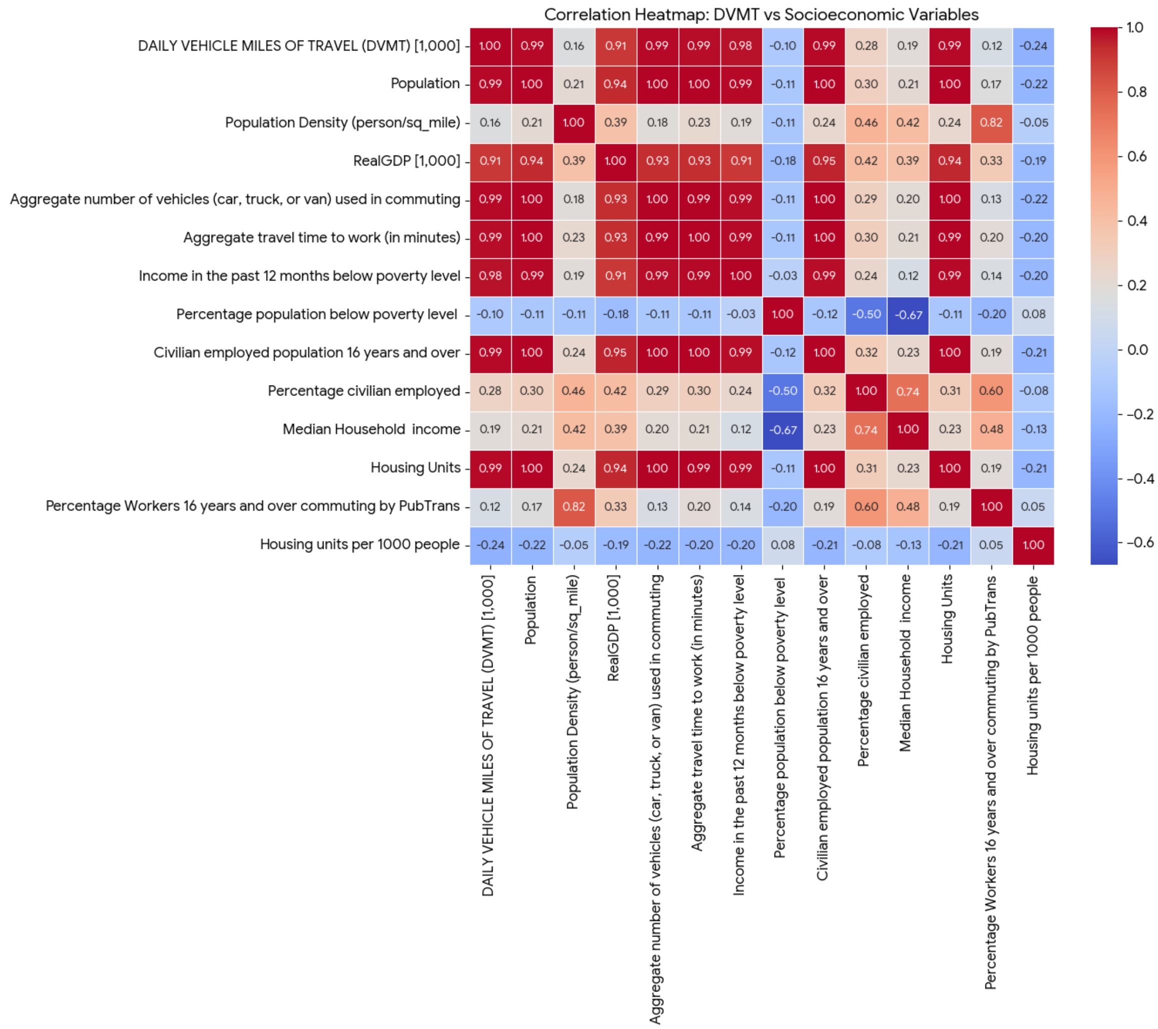
As shown in Figure 1, the period from 2019 to 2023 represents a clear structural break in the historical trajectory of California's transportation dynamics. Overall, the average county-level DVMT decreased by approximately 9% between 2019 and 2023. This statewide contraction indicates a sustained reduction in vehicle travel following the COVID-19 pandemic. In 2020, DVMT declined by 14.86% across nearly all counties, attributable to pandemic lockdowns and reduced commuting. While there was a rebound in 2021 and 2022, most urban counties had not returned to their 2019 DVMT levels by 2023. In addition, the 2023 data show an estimated statewide increase in DVMT of 0.81% relative to 2022. However, the deceleration in the rate of growth—from a 3.96% recovery in 2021, to 1.42% in 2022, and now to 0.81% in 2023—is a critical observation. This slowing momentum suggests that the initial, sharp "rebound" effect from the pandemic's disruption has largely run its course. The state appears to be settling into a new, more stable equilibrium of travel behavior.

Figure 1. Average County-Level DVMT in California (2019–2023)



Correlation analysis reveals clear relationships between changes in DVMT and socio-economic factors. The heatmap (Figure 2) visualizes the correlation matrix. The values range from **-1.0** (perfect negative correlation) to **1.0** (perfect positive correlation).

Figure 2. Correlation Heatmap: DVMT vs Thirteen Socioeconomic Variables



DVMT is primarily driven by the scale of the county (total population, total jobs, total housing) rather than the intensity or efficiency metrics (density, median income, transit usage rates). The variables can be classified into three categories based on their correlation with DVMT:

3.2.1 Strong Positive Correlation (Correlation > 0.90)

These variables track the absolute size of the county. As these numbers increase, DVMT increases almost linearly. This indicates that total travel volume is a direct function of total activity.

- **Aggregate Vehicles Used in Commuting (0.99):** This is the strongest predictor of DVMT because it directly reflects both the volume of travel demand and the number of vehicles generating miles.

- **Population (0.99):** Population is the fundamental driver of travel demand.

- **Housing Units (0.99) & Civilian Employed Population (0.99):** These variables represent the proxies for county size.

- **Aggregate Travel Time to Work (0.99):** Higher total driving volume results in more total time spent on the road.

- **Count of People in Poverty (0.98):** The strong correlation reflects the fact that more populous counties tend to have more people in poverty in absolute terms.

- **Real GDP (0.91):** Economic output is strongly linked to the movement of goods and people.

3.2.2 Weak Positive Correlation (0.00 < Correlation < 0.30)

These variables show a slight positive trend but are not strong predictors of total DVMT.

- **Percentage Civilian Employed (0.28):** Higher employment rates contribute slightly to more driving, but the effect is minor compared to population size.

- **Median Household Income (0.19):** Wealthier areas drive slightly more, but the correlation is weak.

- **Population Density (0.16):** Notably, density has a very weak correlation with total DVMT. While density may reduce per capita driving, dense counties (such as San Francisco or Alameda) still generate substantial total DVMT due to their sheer size.

- **Percentage Commuting by Public Transit (0.12):** This positive correlation is counterintuitive (transit usually reduces driving). However, it indicates that counties with the most transit infrastructure (major urban centers) also have the highest populations and thus the highest DVMT.

3.2.3 Negative Correlation (Correlation < 0.00)

These variables move in the opposite direction of DVMT, though the relationships are relatively weak.

- **Percentage Population Below Poverty Level (-0.10):** A very weak negative link, suggesting that as the *rate* of poverty increases, total DVMT might slightly decrease, likely due to lower vehicle ownership or economic activity.

- **Housing Units per 1000 people (-0.24):** This metric serves as a powerful proxy for the urban-rural divide in California, effectively separating dense population centers from sparsely populated or vacation regions. A low ratio (urban) implies high population density, resulting in massive total DVMT simply due to the sheer number of drivers, even though efficient transit and proximity keep per capita DVMT low. Conversely, a high ratio (rural) implies a small population with many vacation homes; this results in low total DVMT overall, but high per capita DVMT because residents are forced to drive long distances for basic services without transit alternatives.

The geographic pattern of DVMT change highlights the uneven impacts of the pandemic and subsequent behavioral shifts (Table 2 and Figure 3). The key trends are summarized below:

A. The Tech-Hub Decline (Dark and Light Green Zones)

The most significant declines in vehicle travel (-35% to -16%) are heavily concentrated in the San Francisco Bay Area. Counties such as San Francisco, San Mateo, Marin, and Santa Clara show the steepest drops. This may reflect the high prevalence of remote work in the tech sector as well as the dense transit network. Both have drastically reduced commuting into urban cores such as downtown San Francisco. Two rural counties, Mariposa and Sierra, also stand out in dark green. Mariposa's economy is heavily skewed toward tourism. The drop in DVMT here is largely due to policy changes restricting access to Yosemite National Park. Sierra is one of California's least populated counties. The decline in DVMT is attributable to its aging population and net out-migration.

B. Southern and Central California's "Middle Ground" (Yellow Zones)

Southern California shows a much stronger recovery in driving compared to the Bay Area. This stability is likely driven by the logistics/warehousing boom in the Inland Empire (truck traffic) and by a workforce that may have fewer remote work options than in Silicon Valley. Los Angeles, Orange, and San Diego Counties are all in the "yellow" category (-14% to -5%), indicating a slight decrease but not nearly as drastic as the north. On the other hand, the majority of the Central Valley (from Kern up to Butte) is also colored yellow. As an agricultural powerhouse, the Central Valley's economy relies heavily on on-site work, which likely kept DVMT relatively stable compared to the coastal professional hubs.

C. The Resilience of the Logistics Hubs (Light Blue Zones)

The Inland Empire (Riverside and San Bernardino counties) and parts of the Central Valley show much smaller declines (-4.0% to -0.2%, shown in light blue) compared to coastal urban centers. This region is the freight hub of the West Coast. The explosion in e-commerce since 2020 has kept truck traffic high, offsetting reductions in commuter car traffic. Another factor is the housing sprawl. These areas continue to see population growth as people move inland for affordability, sustaining high VMT levels.

D. The Rural and Recreational Rebound (Dark Blue Zones)

In sharp contrast to the coast, the rural periphery of the state shows stability or actual growth in DVMT (0% to 12%, shown in dark blue). Counties such as Modoc, Siskiyou, Alpine, Mono, and Inyo are bucking the statewide downward trend. This pattern may be attributed to several factors:

- Very low population density; a small change in absolute miles can create a big percentage swing.
- Heavy reliance on intercity highways (U.S. 101, I-5, rural 2-lane highways) for essential trips—little ability to substitute transit, walking, or biking.
- Increased recreation and second-home travel into rural and mountain regions after the pandemic, which may have boosted visitor miles.
- Growth in goods movement (trucking along I-5 and east-west freight routes) as e-commerce expanded.

Figure 3. DVMT Changes Across California Counties (2019–2023)

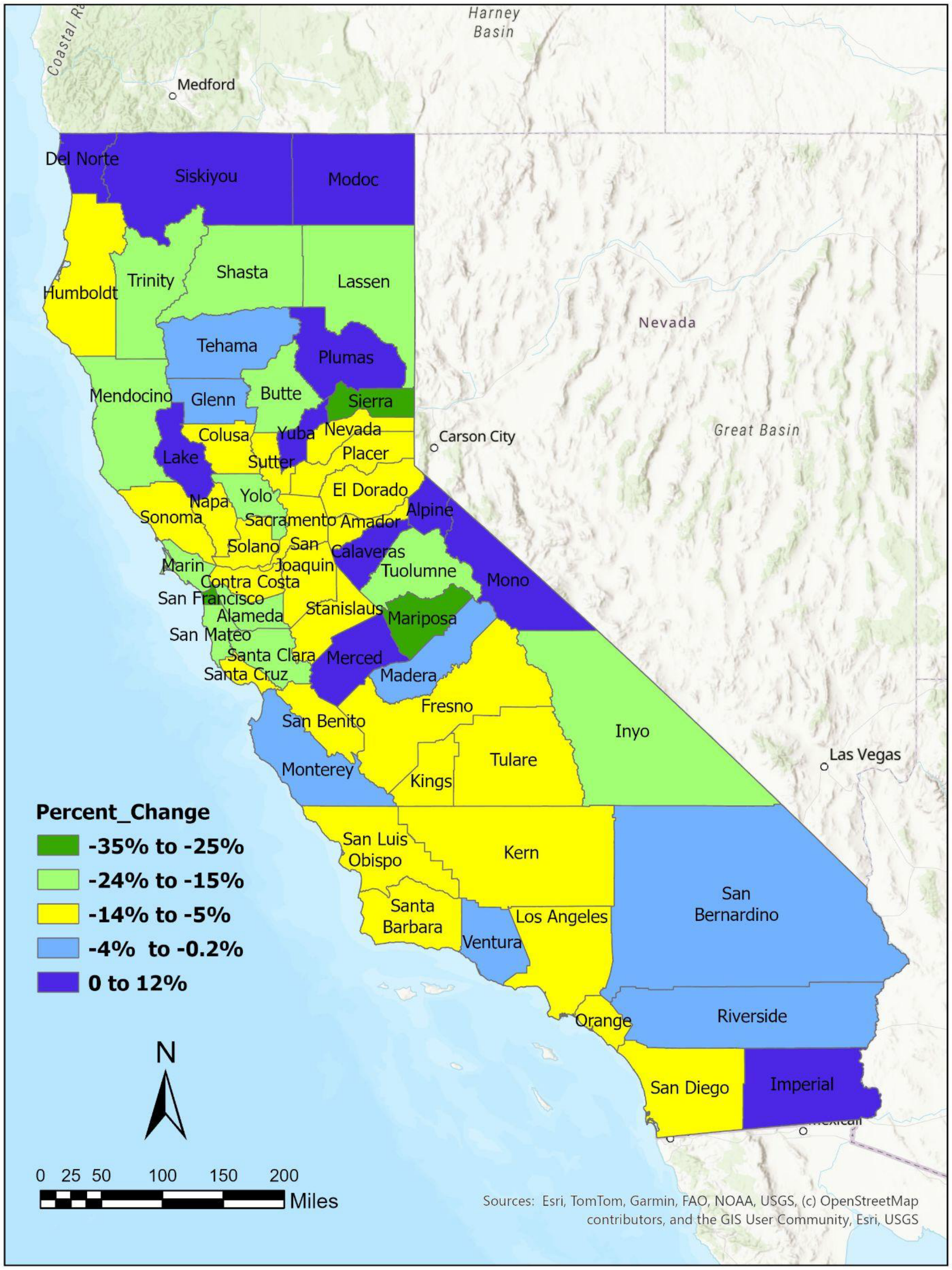


Table 2. California Counties with the Largest Declines and Increases in DVMT (2019–2023)

Largest Declines	Largest Increases
Sierra (-35.0%), San Francisco (-31.8%), Mariposa (-28.3%), Marin (-23.8%), Inyo (-21.3%), San Mateo (-20.0%), Santa Clara (-19.5%), Butte (-18.7%), Trinity (-18.3%), Alameda (-18.2%)	Alpine (+11.3%), Lake (+9.5%), Siskiyou (+7.9%), Calaveras (+3.1%), Modoc (+3.1%), Mono (+2.8%), Plumas (+2.2%), Merced (+1.5%), Yuba (+0.8%), Imperial (+0.5%)

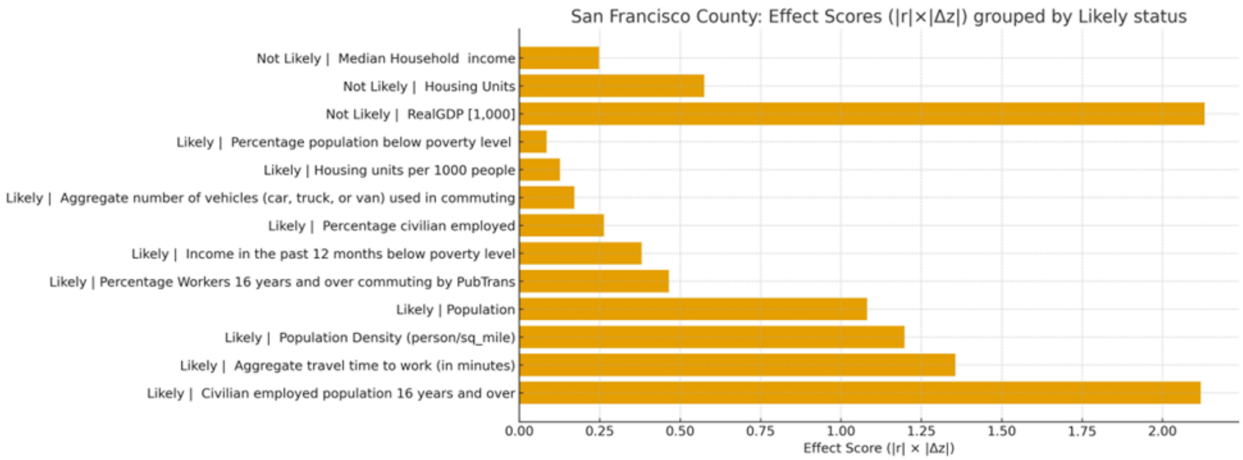
To identify the primary drivers of DVMT change, statewide Pearson correlations, r (DVMT, variable), were computed by pooling 2019 and 2023. For each variable in each county, Δ = 2023−2019 (i.e., absolute percentage change of DVMT between 2019 and 2023) was standardized across counties to Δz. Then we can compute:

$$\textit{Effect score} = |r(DVMT, variable)| \times |\Delta z|$$

The expected DVMT effect indicates whether a change in a given variable would typically push DVMT up or down, determined by the sign of r and Δ (shown in the variable's original units). Variables whose local shifts align with the county's observed DVMT change are identified as likely contributors. Variables with the highest effect scores are the top drivers. Note that correlation is not causation. Local factors (tourism, wildfire recovery, network changes, EV adoption, and telework policy) may also play roles.

For example, San Francisco experienced a 31.8% decline in DVMT between 2019 and 2023. Figure 4 displays the effect scores of the thirteen exploratory variables, grouped by likelihood status. The top three drivers are the variables with the "likely" status (i.e., the same downward trend as San Francisco's DVMT) and high effect scores: civilian employed population 16 years and over, aggregate travel time to work, and population density.

Figure 4. San Francisco: Comparison of Expected DVMT Effect and Likely Status for Exploratory Variables



The prevailing trend is a significant reduction in total commute time—a variable strongly correlated with DVMT—reflecting enduring post-2020 shifts in commute intensity. While a rebound in GDP and relative transit share mitigated this decline, they did not fully offset it.

3.3 Time-Series Cluster Analysis of DVMT and Per Capita DVMT

3.3.1 The K-Means Clustering Method

To further investigate the spatial-temporal dynamics of vehicle travel across California counties, we used a time-series cluster-analytic approach to examine 2019–2023 DVMT and per capita DVMT trajectories. Specifically, we applied the K-means clustering algorithm with k = 5 in ArcGIS Pro 3.6 and ChatGPT (GPT-5.2 Pro) to analyze standardized measures (z-scores) of DVMT and per capita DVMT. By grouping counties with similar travel profiles and then examining their temporal dynamics, we aim to: (1) characterize how total and per capita driving changed before (2019), during (2020–2021), and after (2022–2023) the acute pandemic period in different types of counties; and (2) derive cluster-specific policy implications for transportation planning and practice.

The analysis follows a structured workflow consisting of data preparation, feature standardization, time-series feature extraction, and unsupervised clustering with the K-means algorithm. Each county's DVMT and per capita DVMT trajectories over 2019–2023 were combined into a multivariate feature vector. Standardization was applied to compute the year-wise z scores to prevent magnitude differences from dominating clustering outcomes:

$$z_i = (x_i - \mu) / \boldsymbol{\sigma}$$

Where z_i is the DVMT z score or per capita DVMT z score for a county, x_i is the raw DVMT or per capita DVMT value for a county, μ is the mean across all counties for that year, and σ is the standard deviation.

Below is the objective function of the K-means clustering algorithm:

$$J = \Sigma_k \Sigma_i ||x_i - \mu_k||^2$$

where μ_k is the centroid of cluster **k**, and x_i represents the feature vector of county **i**. The algorithm minimizes within-cluster variance by iteratively updating cluster assignments and centroids. We experimented with different groupings and selected k = 5 based on the silhouette coefficient, a measure of the normalized difference between intra- and inter-cluster distances. The optimal k balances compactness and separation. The resulting classification (Figure 5) demonstrates a highly robust structure and strong cluster separation:

1. Inertia Decomposition: The analysis attributes 91.88% of the total inertia (580.000) to the between-cluster variance, leaving only 8.12% within clusters. This indicates that the chosen five clusters capture fundamental distinctions in county travel patterns.

2. Convergence: The optimal solution was achieved after 11 iterations, reaching a low final within-cluster variance of 0.889.

3. ANOVA Contribution: Analysis of variance (ANOVA) confirms that the classification significantly explains the variance for all ten input variables. F-statistics range from 101.026 (2019 per capita DVMT Z) to 176.219 (2022 DVMT Z), with all tests achieving high significance ($P < 0.0001$).

4. Silhouette Measure[1]: The overall mean silhouette width is 0.62, suggesting a reasonably good fit, where objects are well-matched to their own cluster. The largest cluster (Cluster 3) contains 42 of the 58 counties, representing the "average" or typical county profile. It also maintains the strongest internal cohesion (Mean Silhouette Score of 0.71) in the k = 5 solution.

[1] The **Silhouette Score** is a metric used to evaluate the quality of K-means clustering algorithm. It measures how similar a data point is to its own cluster compared to other clusters. A silhouette score closer to 1 indicates that the data point is well-matched to its own cluster and far from neighboring clusters.

Figure 5. Silhouette Scores

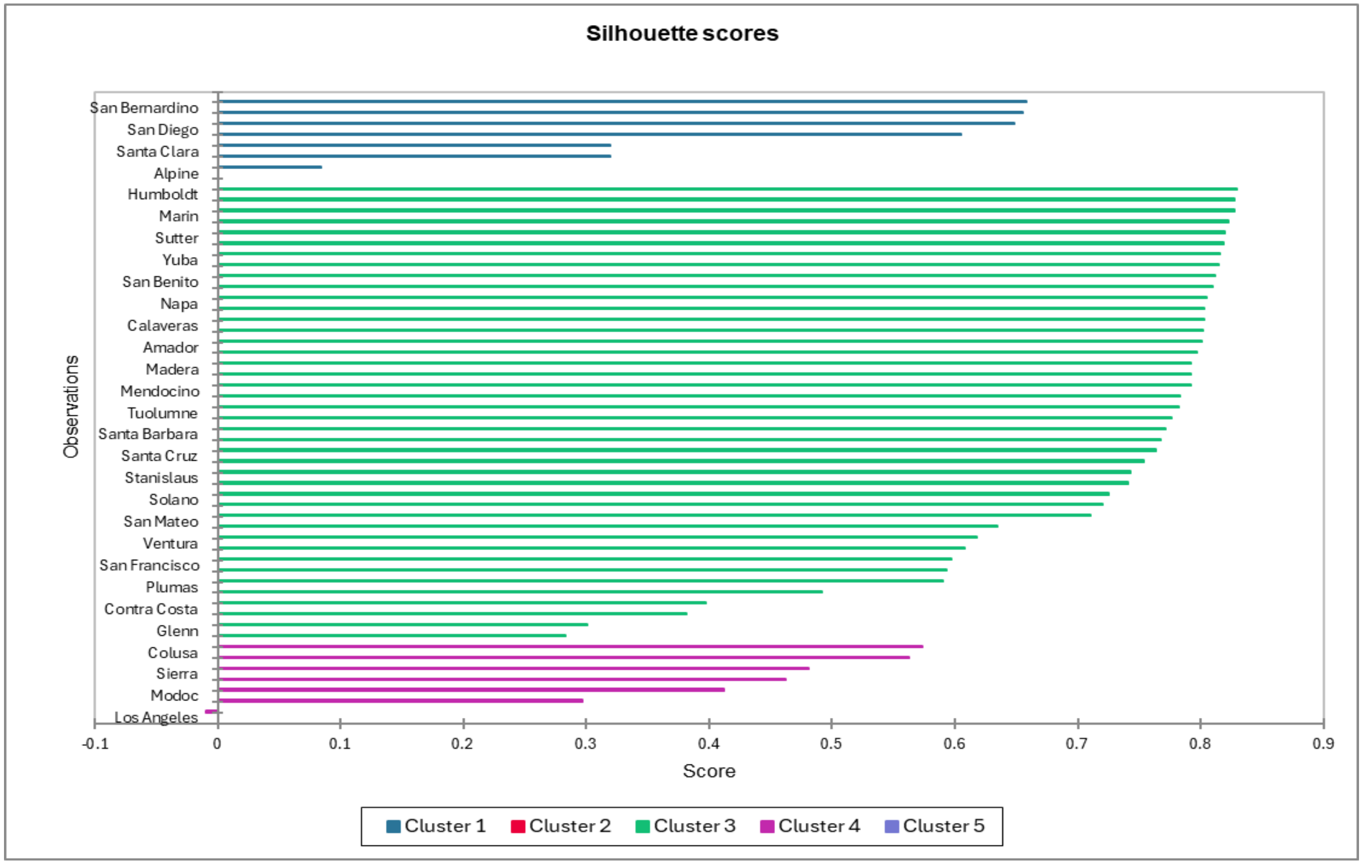


Figure 6. Results of the K-Means Cluster Analysis (k = 5)

The five clusters (Figure 6) delineate distinct profiles based on the magnitude (DVMT Z-score) and intensity (per capita DVMT Z-score) of travel, particularly revealing differential responses to the pre-pandemic (2019), middle/peak impact (2020–2021), and post-pandemic recovery (2022–2023) periods. The clusters can be interpreted as:

- Cluster 1 (Major Urban/Suburban). This cluster includes seven counties: Alameda, Santa Clara, Sacramento, Orange, San Bernardino, San Diego, and Riverside. They are the economic engines with extensive commuting corridors, with high total DVMT and moderate per capita DVMT.

- Cluster 2 (Alpine Outlier). Alpine County is a small rural county with negligible total DVMT at the statewide level but an extremely high per capita DVMT, more than 5 standard deviations above the mean.

- Cluster 3 (Mixed/Metro Core). This is the largest cluster of 42 mid-sized metropolitan, suburban, and rural counties, including San Francisco, San Mateo, Contra Costa, Kern, Fresno, San Joaquin, etc. These counties exhibit moderate total DVMT and generally lower per capita DVMT (especially San Francisco) due to higher density and greater transit options.

- Cluster 4 (Rural/Recreational). This cluster contains seven counties: Sierra, Inyo, Mono, Modoc, Siskiyou, Colusa, and Trinity. They are sparsely populated rural counties with relatively low total DVMT but very high per capita VMT, driven by tourism, long distances between service locations, and a lack of transit.

- Cluster 5 (Mega-Urban). Los Angeles County is another unique outlier. It has a relatively low per capita DVMT, but its immense scale (nearly 200 million DVMT in 2023) is over 6 standard deviations above the mean. This highlights that its 23% contribution to statewide DVMT is unique and incomparable to other regions.

3.3.2 Temporal Patterns in Cluster-Mean DVMT

Figure 7 illustrates the variation in average DVMT z scores across clusters and time periods. In general, DVMT declined substantially in 2020, partially rebounded through 2022 and 2023 but remained below 2019 levels at the end of the study period. Cluster 2 is not included here as it is an outlier with extremely small DVMT values. As shown in Figure 8, the mid-pandemic DVMT indices for large metropolitan (Cluster 1) and typical counties (Cluster 3) are approximately 87 and 86, respectively (2019 = 100). Post-period indices rise to around 89 for these clusters. High per capita rural counties (Cluster 4) also experience a large initial drop, with a mid-period index near 82 and a post-period index of about 89. Los Angeles County (Cluster 5) recovers somewhat more strongly, reaching a post-period index close to 93. Alpine County (Cluster 2) is an exceptional case. It is the least populous county in California, with roughly 1200 residents. Because the baseline traffic is so incredibly low, it takes very few additional vehicles to create a "double-digit" percentage spike.

Figure 7. Cluster-Average DVMT (2019–2023)

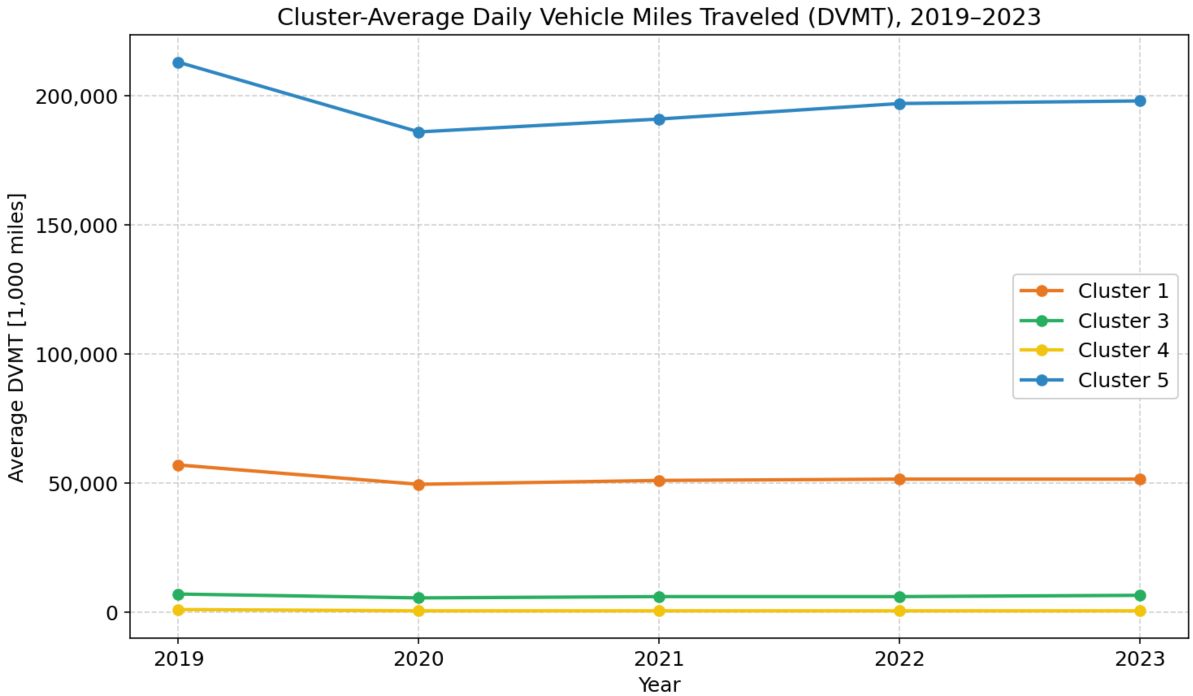
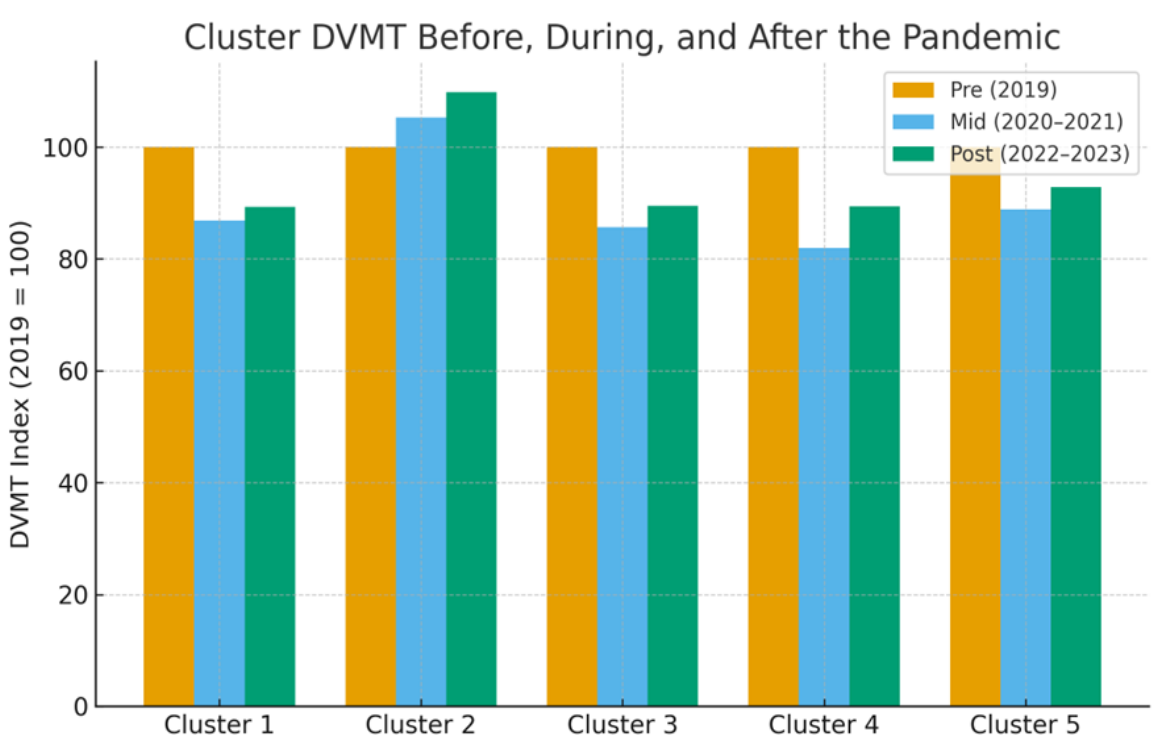


Figure 8. Cluster DVMT Indices (2019 = 100) by Pre-, Mid-, and Post-Pandemic Phases



3.3.3 Temporal Patterns in Cluster-Mean Per Capita DVMT

Per capita DVMT shows even more persistent declines (Figure 9). The cluster index chart (Figure 10) shows that large metropolitan and typical counties stabilize with post-period per capita indices around 88–89, implying that average per-person driving remains roughly 10–12 percent below 2019 levels. Rural counties with high per capita DVMT exhibit the largest lasting reductions, with a post-period index near 82 (about 18 percent below the 2019 baseline). Los Angeles County is closest to full recovery, with a post-period index of approximately 95, but it remains slightly below its pre-pandemic level.

Figure 9. Cluster-Average Per Capita DVMT (2019–2023)

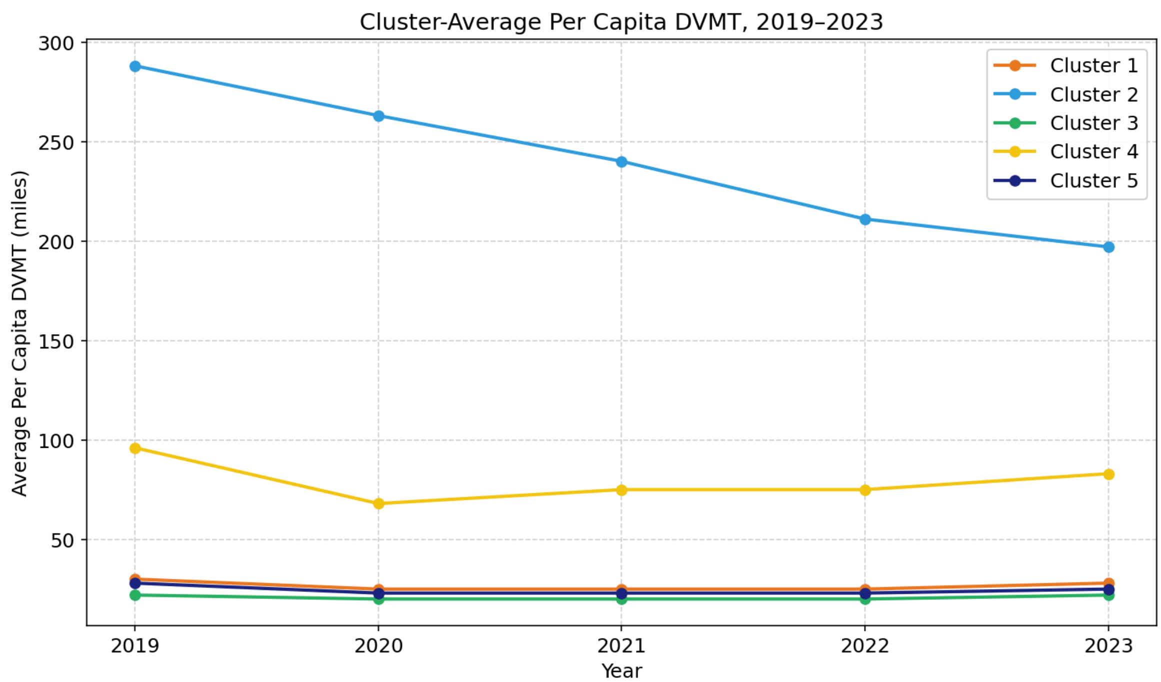
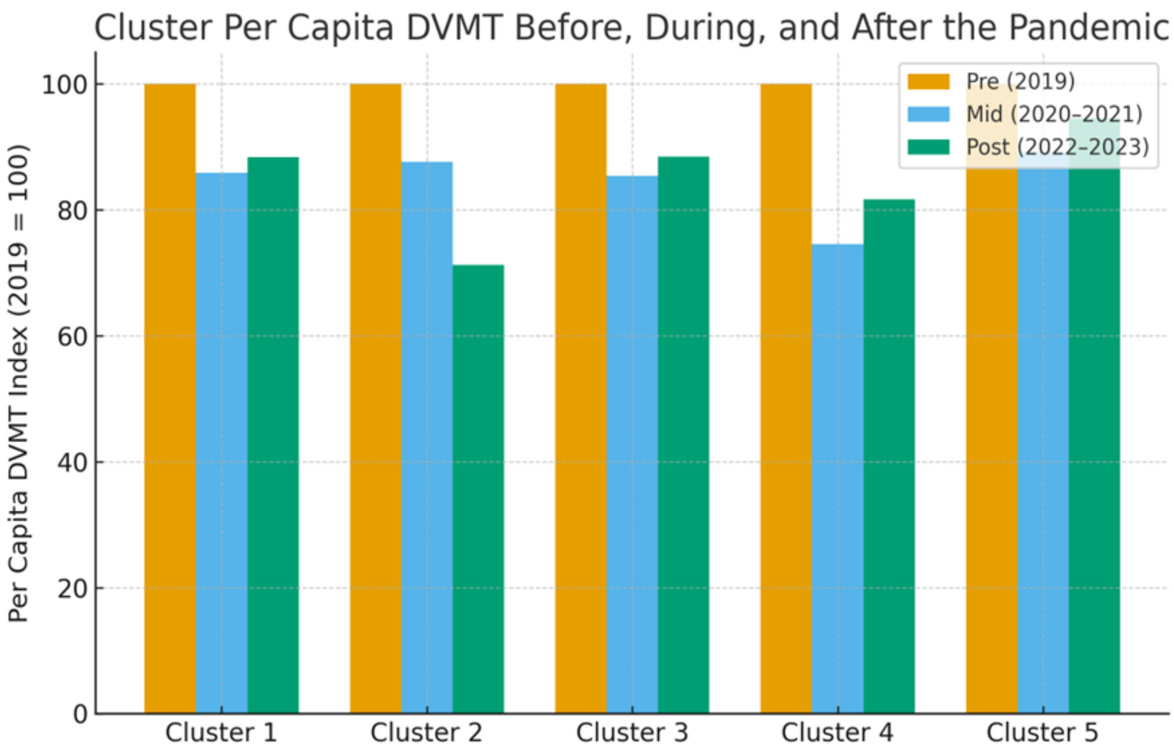


Figure 10. Cluster Per Capita DVMT Indices (2019 = 100) by Pre-, Mid-, and Post-Pandemic Phases

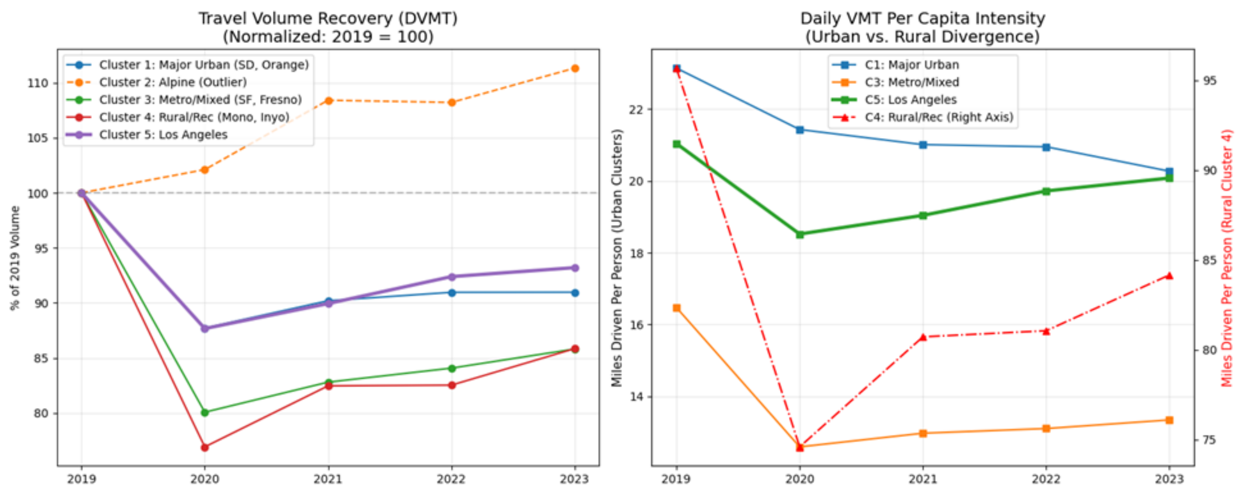


The cluster-based results underscore the heterogeneity of these changes. Metropolitan clusters exhibit relatively strong DVMT recovery but maintain modest per capita reductions, consistent with durable telecommuting. Typical counties mirror these trajectories at lower absolute volumes. Rural counties with high per capita DVMT experience the largest reductions, which may reflect shifts in long-distance or recreational travel as well as demographic change.

3.4 Implications for Transportation Policy and Practice

Figure 11 illustrates a pronounced divergence between the "Rural Recreational" cluster (Cluster 4) and the "Urban Jobs" clusters (Clusters 1 and 5). All clusters experienced a synchronized collapse in DVMT during the early pandemic, but the recovery was highly uneven—muted and incomplete in dense "Urban Jobs" areas, almost full in typical rural counties, and above baseline in rural recreational destinations. Over time, the gap between high-VMT rural recreational counties and lower-VMT urban counties widens, underscoring the inadequacy of uniform, statewide approaches to VMT reduction. Instead, the observed spatial-temporal patterns call for differentiated, context-sensitive policy responses.

Figure 11. The Divergence Between "Rural Recreational" (Cluster 4) and "Urban Jobs" (Clusters 1 & 5)



3.4.1 For Urban & Suburban Clusters (1, 3, and 5)

In urban and suburban counties, per capita commuting DVMT remains persistently below pre-pandemic levels, suggesting that peak-period travel demand has flattened. This shift is consistent with the widespread adoption of remote and hybrid work arrangements. For transit agencies, these trends imply that traditional 9-to-5, peak-focused service models may no longer be optimal. Rather, agencies may need to reorient toward all-day, frequency-based networks that prioritize accessibility for non-work trips, including shopping, healthcare, and social or recreational travel.

Lower per capita DVMT in large, high-income counties with strong tax bases (e.g., Los Angeles, Santa Clara) raises concerns about the long-term sustainability of transportation revenue streams that depend on fuel consumption. Because fuel use appears to be declining faster than DVMT in these regions—partly due to rapid adoption of electric and more fuel-efficient vehicles—gas tax receipts are likely to erode. These dynamics reinforce the importance of piloting and evaluating Road Usage Charging (RUC) schemes to better align revenue with system use in a decarbonizing vehicle fleet.

3.4.2 For Rural and Recreational Clusters (2, 4)

Residents in the “Rural Recreational” cluster routinely travel more than 60 to 80 miles per person per day, reflecting long distances to essential services, employment, and recreation. Under these conditions, a uniform per-mile VMT charge would impose a substantially higher financial burden on rural households, not because they choose to travel more, but because their built environment and limited service accessibility require them to do so. In this way, one-size-fits-all mileage pricing exemplifies a per capita equity trap: a policy that appears neutral in rate design but amplifies existing spatial and socioeconomic inequities by pricing unavoidable travel at the same level as optional or substitutable travel. Transportation policies must differentiate between "luxury" or tourist VMT and "essential" resident VMT. Much of the VMT in Cluster 4 counties is attributable to tourism and recreation rather than to daily resident travel. The resulting surge in vehicle activity accelerates wear and tear on rural roadways, even though local governments often have relatively small tax bases. State-level funding formulas may therefore need to incorporate measures of "visitor VMT" alongside resident population, so that jurisdictions hosting high volumes of recreational travel are not structurally underfunded.

3.4.3 General Policy (SB 743 & VMT Reduction)

California's shift to VMT as a primary transportation impact metric under CEQA (SB 743) interacts unevenly with the spatial patterns highlighted in Figure 11. In rural counties such as Inyo and Mono, new housing development will, by definition, generate relatively high per capita VMT due to sparse land use patterns and limited mode choices. Applying uniform VMT thresholds in these contexts risks effectively discouraging or blocking the needed housing development in already-constrained rural markets.

To reconcile equity, housing, and climate objectives, regulatory frameworks may require exemptions or modified VMT thresholds for rural housing projects. At the same time, VMT reduction efforts can be more aggressively pursued in urban and suburban clusters (1 and 5), where higher densities and more extensive transit networks make substantial reductions in VMT both technically feasible and socially acceptable.

In sum, the comprehensive dataset covering California's fifty-eight counties reveals that the statewide aggregate stability of VMT masks a profound regional and socioeconomic divergence. Overall, the simultaneous decline in total DVMT and per capita DVMT, followed by stabilization below 2019 values, suggests that the COVID-19 shock has produced partially structural changes in travel behavior rather than a purely cyclical disturbance. Urban and high-income regions saw the most pronounced declines due to the adoption of remote work and continued hybrid arrangements, while rural and sparsely populated areas were more resilient. These findings have implications for transportation planning, environmental regulation, and infrastructure investment, particularly in addressing post-pandemic travel demand. They highlight the need for policy instruments that (1) account for the differentiated roles of commuting versus non-work travel, (2) protect rural and recreational communities from regressive impacts of flat VMT-based charges, and (3) strategically focus VMT reduction efforts where alternative modes and land use configurations make them most effective. In this sense, the post-pandemic period represents not only a challenge, but also an opportunity to align transportation, equity, and climate objectives in more spatially nuanced ways.

4. Linear Regression and Tree-Based Modeling of Per Capita DVMT

While total DVMT often rises simply because a region is adding residents, per capita DVMT serves as a pure efficiency metric, revealing how individuals are responding to economic changes, land use policies, and transportation investments. Accordingly, this study analyzes per capita DVMT, given its ability to isolate genuine shifts in travel behavior from the overwhelming "noise" of population growth. By focusing on the per capita trend, this approach helps assess the "Decoupling Hypothesis"—the economic theory that GDP growth can be decoupled from VMT growth (Tapio, 2005; Millard-Ball & Schipper, 2010; Garceau et al., 2014). It can also assist planners to detect critical structural breaks—such as the "saturation" of driving demand or the success of transit-oriented development—that remain invisible in aggregate data. Ultimately, this distinction is vital for accurate forecasting and climate policy; it ensures that infrastructure planning is based on actual human demand rather than a simple extrapolation of demographic trends, preventing the costly overbuilding of highway capacity.

4.1 Selection of Independent Variables

Since we chose per capita DVMT as the dependent variable, we also selected and recalculated nine rate-based socioeconomic measures as independent variables. Moreover, we used the Variance Inflation Factor (VIF) to detect and address multicollinearity (Table 3). For each independent variable $\mathbf{x_i}$, VIF is calculated as:

$$\mathbf{VIF_i = 1 / (1 -} R_i^2\mathbf{)}$$

where $\mathbf{R_i^2}$ is the coefficient of determination from regressing xi on all other predictors. A VIF of 1 indicates no correlation; values between 1 and 5 suggest moderate correlation (acceptable), and values above 10 indicate high multicollinearity and potential redundancy.

The initial VIF analysis identified several variables with VIFs exceeding 10, indicating significant multicollinearity.

Table 3. Initial Results of the VIF Analysis

Variable	VIF
Population Density (person/sq_mile)	4.27
Per Capita Real GDP	16.08
Housing Units per 1000 people	7.52
Percentage of population below poverty level	16.81
Aggregate travel time to work (in minutes) per employee	41.98
Aggregate number of vehicles (car, truck, or van) used in commuting per employee	104.33
Percentage of civilians employed	129.82
Median household income	61.74
Percentage of workers 16 years and over commuting by public transportation	6.4

Several variables showed strong correlations ($|r| > 0.7$), for example:

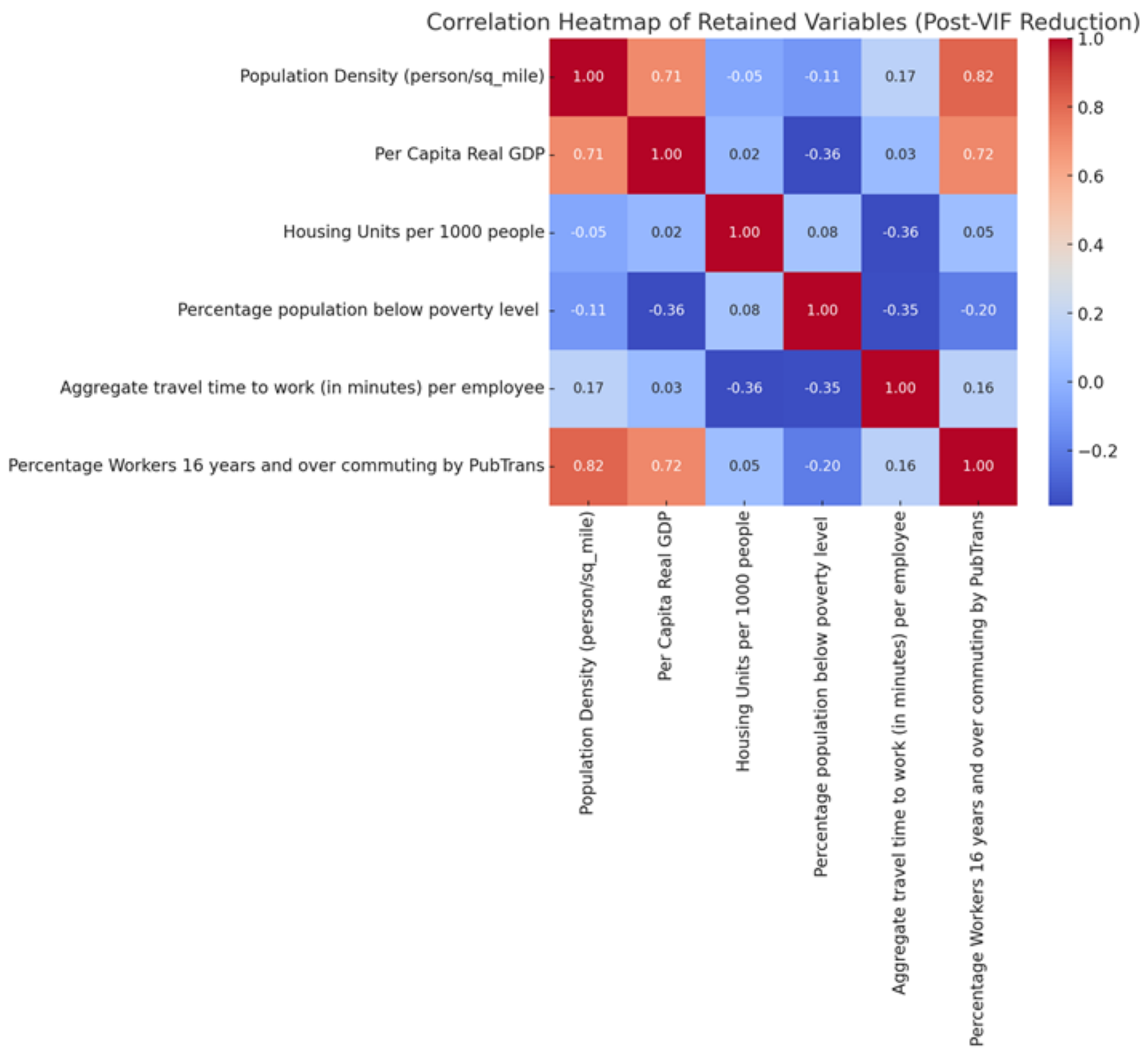
- Population density ↔ Public transit use (0.82)
- Per capita real GDP ↔ Median household income (0.75)
- Per capita real GDP ↔ Public transit use (0.72)

Strong negative relationships also existed, such as:

- Aggregate vehicles ↔ Public transit use (–0.77)
- Per capita real GDP ↔ Aggregate vehicles (–0.67)

A stepwise reduction approach was applied, in which variables with the highest VIF values above the threshold were iteratively removed until all remaining variables had VIF values below 10. As a result, three variables—"Percentage of civilian employed," "Aggregate number of vehicles per employee," and "Median household income"—were removed. A correlation heatmap was generated to visualize pairwise relationships among the retained variables (Figure 12). The heatmap confirms that all remaining predictors show moderate intercorrelations, ensuring minimal multicollinearity. The refined variable set "Population density," "Per capita real GDP," "Aggregate travel time per employee," "Housing units per 1000 people," "Percentage of population below poverty level," and "Percentage of workers 16 years and over commuting by public transportation"—can be confidently used in subsequent regression modeling without undue collinearity bias.

Figure 12. Correlation Heatmap: Six Retained Independent Variables for Per Capita DVMT Models



4.2 Comparison of Linear and Tree-Based Regression Models

Using the 2019–2023 California county dataset with the six post-VIF variables (Figure 12), we developed and compared six regression models—including Ordinary Least Squares (OLS), Ridge, Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, Random Forest, and Gradient Boosting. These models were chosen to represent both linear and nonlinear predictive frameworks. The objectives are to (1) identify the key drivers of per capita DVMT at the county level, (2) assess the capacity, performance, and interpretability of linear and tree-based models to capture spatial variability across California's diverse regions, and (3) derive insights that inform policy and planning.

The key difference between these models lies in how they learn: linear models (OLS, Ridge, LASSO, Elastic Net) fit a straight line or plane to the data, whereas tree ensembles (Random Forest, Gradient Boosting) capture complex, nonlinear patterns by combining multiple decision trees.

4.2.1 Common Linear Modeling Template

Let $\mathbf{y}$ be the dependent variable and $\mathbf{x_1, x_2, x_3, x_4, x_5, x_6}$ be the six independent variables (Figure 12). For each county **i (i = 1 to 58)**, we have:

$$\mathbf{\hat{y}_i = b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} + b_5x_{i5} + b_6x_{i6}}$$

Here, $\mathbf{b_0}$ is the intercept and $\mathbf{b_1, \ldots, b_6}$ are the coefficients for the six predictors. The error for observation i is:

$$\mathbf{Error_i = y_i - \hat{y}_i}$$

The total squared error over all **n** (**n=58**) observations is:

$$\mathbf{Total\ squared\ error = \Sigma_i\,(y_i - \hat{y}_i)^2 = \Sigma_i\,Error_i^2}$$

Each linear method (OLS, Ridge, LASSO, Elastic Net) chooses $\mathbf{b_0, \ldots, b_6}$ to minimize: Total squared error + Penalty. However, the penalty term differs across models.

4.2.2 OLS (Ordinary Least Squares)

OLS has no penalty term. It chooses $b_0, \ldots, b_6$ to make the total squared error as small as possible.

- Model equation:

$$\mathbf{\hat{y}_i = b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} + b_5x_{i5} + b_6x_{i6}}$$

- Objective (no penalty):

Minimize over $\mathbf{b_0,\ldots,b_6}$: $\mathbf{\Sigma_i\,(y_i - \hat{y}_i)^2}$
$\mathbf{= \Sigma_i\,[y_i - (b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} + b_5x_{i5} + b_6x_{i6})]^2}$

For OLS, coefficients are easy to interpret but can be unstable when predictors are correlated or when the sample size is small.

4.2.3 Ridge Regression

Ridge Regression retains the OLS objective but adds an L2 penalty on the six coefficients $\mathbf{b_1, \ldots, b_6}$.

- Model equation (same as OLS):

  $\hat{y}_i = b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} + b_5x_{i5} + b_6x_{i6}$

- Penalty term (L2):

  Penalty_Ridge $= \lambda(b_1^2 + b_2^2 + b_3^2 + b_4^2 + b_5^2 + b_6^2)$.

- Objective with penalty:

  Minimize over $b_0,\ldots,b_6$: $\Sigma_i\,(y_i - \hat{y}_i)^2 + \lambda(b_1^2 + b_2^2 + b_3^2 + b_4^2 + b_5^2 + b_6^2)$

For Ridge, $\lambda \geq 0$ controls how strongly large coefficients are penalized. Coefficients are shrunk toward zero but rarely become exactly zero. It helps when the six predictors are correlated or when controlling overfitting is important.

4.2.4 LASSO (Least Absolute Shrinkage and Selection Operator)

LASSO uses an L1 penalty on the absolute values of the six coefficients $b_1, \ldots, b_6$, which can set some of them exactly to zero.

- Model equation (same as OLS):

  $\hat{y}_i = b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} + b_5x_{i5} + b_6x_{i6}$

- Penalty term (**L1**):

  Penalty_LASSO $= \lambda(|b_1| + |b_2| + |b_3| + |b_4| + |b_5| + |b_6|)$

- Objective with penalty:

  Minimize over $b_0,\ldots,b_6$: $\Sigma_i\,(y_i - \hat{y}_i)^2 + \lambda(|b_1| + |b_2| + |b_3| + |b_4| + |b_5| + |b_6|)$

For LASSO, $\boldsymbol{\lambda} \geq \mathbf{0}$ controls how strongly nonzero coefficients are penalized. It performs embedded variable selection among the six predictors by setting some coefficients exactly zero.

4.2.5 Elastic Net

Elastic Net regression extends linear modeling by incorporating both **L1 (Lasso)** and **L2 (Ridge)** regularization to control multicollinearity and prevent overfitting (Chung, et al., 2024).

- Model equation (same as OLS):

  $$\hat{y}_i = b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} + b_5x_{i5} + b_6x_{i6}$$

- Penalty term: Elastic Net combines **L1** (LASSO-style) and **L2** (Ridge-style) penalties on the six coefficients **b_1, …, b.**

  $$\text{Penalty\_EN} = \lambda[\alpha(|b_1| + |b_2| + |b_3| + |b_4| + |b_5| + |b_6|) + (1 - \alpha)(b_1^2 + b_2^2 + b_3^2 + b_4^2 + b_5^2 + b_6^2)]$$

- Objective with penalty:

  Minimize over $b_0,\ldots,b_6$: $\Sigma_i (y_i - \hat{y}_i)^2 + \lambda[\alpha\Sigma_j|b_j| + (1 - \alpha)\Sigma_j b_j^2]$, $j = 1,\ldots,6$.

Note that λ controls overall regularization strength:

- **α ($0 \leq \alpha \leq 1$)** mixes the two penalties.
- $\alpha = 1 \rightarrow$ LASSO.
- $\alpha = 0 \rightarrow$ Ridge.

Elastic Net provides stable coefficient estimates when predictors are highly correlated and facilitates variable selection by shrinking less informative coefficients toward zero.

4.2.6 Random Forest

Random Forest is an ensemble learning algorithm that constructs a collection of decision trees and averages their predictions to reduce variance and improve generalization. Each tree is trained on a bootstrapped sample of the six independent variables $x_1,\ldots,x_6$, with a random subset of predictors considered at each split.

Let $x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6})$ be the six predictors for observation i.

- Model formula (regression):

$$\hat{y}_i = (1/M)[T_1(x_i) + T_2(x_i) + \ldots + T_m(x_i)],$$

where $T_m(x_i)$ is the prediction from the m-th tree and M is the number of trees.

Each tree is trained on a bootstrap sample and uses random subsets of the six predictors at each split. This nonparametric approach captures nonlinear relationships and high-order interactions among predictors without explicit functional assumptions. By averaging across trees, it also reduces variance and helps avoid overfitting. Feature importance was assessed based on the mean decrease in impurity across all trees.

4.2.7 Gradient Boosting

Gradient Boosting is another ensemble tree-based method that develops models sequentially. It builds an additive model of many small trees using the six predictors $x_1,\ldots,x_6$, where each new tree is trained to correct the errors of the current model.

Let $x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6})$ be the six predictors for county **i**.

Start with a simple initial prediction, such as the mean of **y**:

$$\hat{y}_0(x_i) = c \text{ (e.g., } c = \text{average of all } y_i).$$

Add **M** small trees sequentially:

$$\hat{y}_m(x_i) = \hat{y}_0(x_i) + \nu T_1(x_i) + \nu T_2(x_i) + \ldots + \nu T_m(x_i),$$

Where ν is a small learning rate and $T_k(x_i)$ is the prediction from the k-th small tree.

Final prediction after M trees:

$$\hat{y}_i = \hat{y}_m(x_i) = c + \nu \Sigma_k T_k(x_i).$$

The algorithm minimizes a differentiable loss function—typically squared error—through gradient descent in function space. Each tree is trained on the residuals (errors) from the previous model, using the six predictors. Gradient Boosting is very flexible and often achieves strong predictive performance on tabular data. However, the interpretation is more complex; feature importance and PDP plots are frequently used.

4.3 Model Results

4.3.1 Model Performance

The six models were estimated using ChatGPT (GPT-5.2 Pro) to predict the same year per capita DVMT for 58 California counties. The 2019–2022 observations were used as training data. The 2023 observations were held out as a test set for out-of-sample evaluation. Model performance is evaluated using four metrics for both the training (2019–2022) and test (2023) sets: the coefficient of determination (R^2), mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE).

Table 4 reports both the in-sample (training) and out-of-sample (test) performance metrics for all six models. Higher R^2 (coefficient of determination) and lower MAE (Mean Absolute Error)/RMSE (Root Mean Square Error)/MAPE (Mean Absolute Percentage Error) indicate a better fit. This analysis focuses on out-of-sample performance, as it reflects how well each model is expected to generalize to new observations.

Table 4. Model Performance Metrics

Model	R^2	MAE	RMSE	MAPE (%)
OLS -Train	0.66	14.02	20.52	41.27
OLS -Test	0.47	15.50	21.63	47.18
Ridge - Train	0.66	14.02	20.52	41.30
Ridge - Test	0.47	15.50	21.63	47.20
LASSO - Train	0.64	13.08	21.08	34.70
LASSO - Test	0.49	13.94	21.10	38.5
Elastic Net - Train	0.64	13.05	21.10	34.58
Elastic Net - Test	0.50	13.91	21.09	38.38
Random Forest - Train	0.98	2.37	4.94	5.69
Random Forest - Test	0.82	4.32	12.63	7.75
Gradient Boosting - Train	0.99	0.52	0.65	1.85
Gradient Boosting - Test	0.91	3.60	8.82	8.17

Figure 13. Predictive Performance of the Six Regression Models

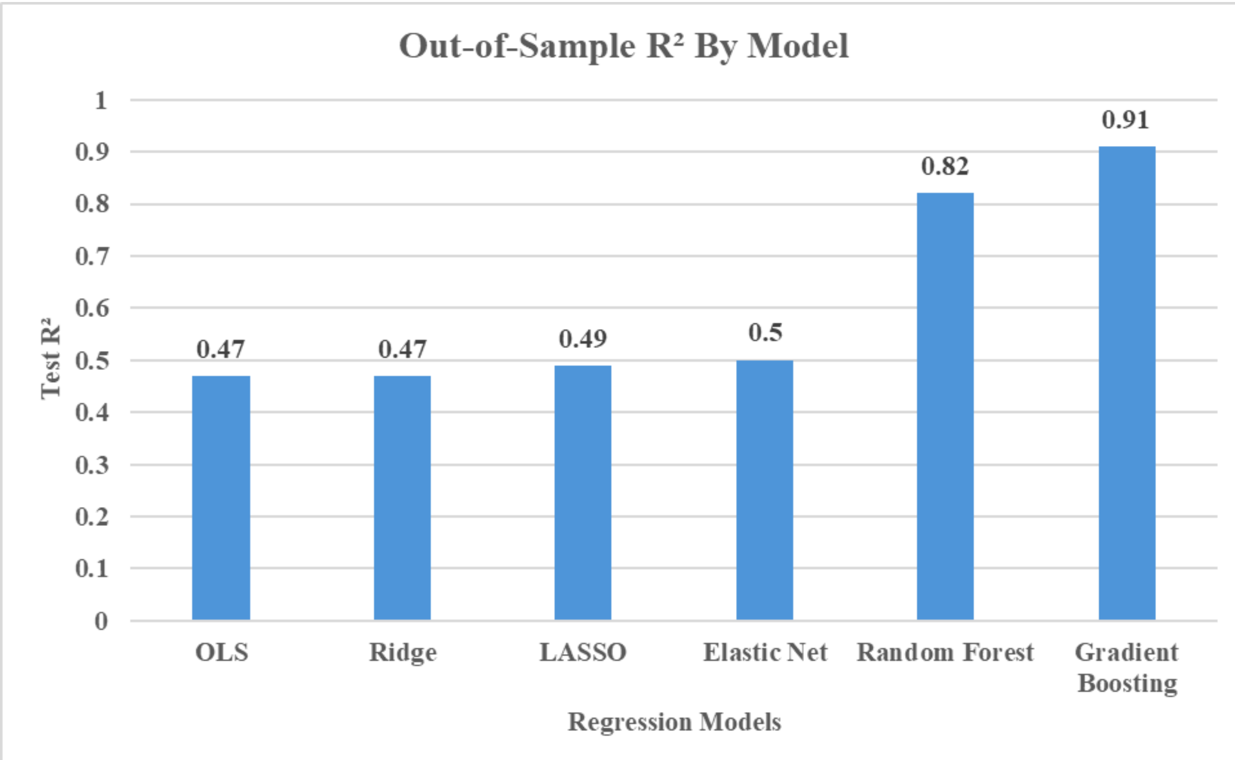


Figure 13 shows that the four linear models (OLS, Ridge, LASSO, and Elastic Net) achieve moderate test-set $\mathbf{R^2}$ values between approximately 0.47 and 0.50. In contrast, the two tree-based ensemble models, Random Forest and Gradient Boosting, perform substantially better, with test R^2 values of approximately 0.82 and 0.91, respectively. This suggests that the outcome variable exhibits nonlinearities and interactions that cannot be fully captured by linear specifications, even when regularization is used.

OLS and Ridge regression deliver similar performance. Both models obtain training $\mathbf{R^2}$ values around 0.60 and test $\mathbf{R^2}$ values around 0.47, with relatively large MAE, RMSE, and MAPE on the test set. The similarity between OLS and Ridge suggests that after applying the VIF selection process, Ridge no longer provides meaningful alteration to the fitted coefficients at the current level of regularization.

LASSO and Elastic Net introduce L1-based sparsity, effectively shrinking some coefficients to zero and performing implicit variable selection. This yields modest but consistent improvements in predictive accuracy. LASSO increases test R^2 to about 0.49 and reduces test MAPE relative to OLS and Ridge, indicating better generalization and lower relative error. Elastic Net performs best among the linear models, with a test $\mathbf{R^2}$ of 0.50 and the lowest test MAE, RMSE, and MAPE in this group. The small gap between its training and test metrics suggests that Elastic Net strikes a favorable balance between bias and variance while retaining interpretability through a linear structure.

The two tree-based ensemble models, Random Forest and Gradient Boosting, achieve dramatically higher in-sample $\mathbf{R^2}$ values (about 0.98–0.99) and much smaller training errors than any linear alternative. These near-perfect fits indicate that the models are highly flexible and capable of capturing complex nonlinear relationships and interactions in the data. Although such flexibility introduces a risk of overfitting, their out-of-sample performance remains strong, with test $\mathbf{R^2}$ values of approximately 0.83 for Random Forest and 0.91 for Gradient Boosting.

In terms of error metrics, both ensembles sharply reduce MAE, RMSE, and MAPE relative to the linear models. Random Forest delivers slightly lower test-set MAPE (7.75%), indicating smaller typical percentage deviations between predictions and observed values. Gradient Boosting, on the other hand, attains slightly higher test $\mathbf{R^2}$ and lower test RMSE (8.82), implying that it explains more of the variance in the outcome and produces fewer large squared errors. Despite visible train–test gaps, both methods substantially improve predictive accuracy over the regularized linear models.

Overall, the results indicate that (a) there is no strong benefit to using simple OLS or (with this tuning) Ridge compared to the regularized sparse models; and (b) Random Forest and Gradient Boosting are the most accurate models for this prediction task, capturing nonlinear structure that linear models cannot. Between them, Gradient Boosting is preferred when maximizing R^2 and minimizing RMSE are the primary objectives, whereas Random Forest may be slightly more attractive if minimizing percentage error (MAPE) or model robustness is prioritized. If interpretability and simple coefficient-based insights are crucial, Elastic Net emerges as the strongest choice among the linear models, offering improved predictive performance relative to OLS, Ridge, and LASSO while retaining a transparent parametric representation. In practice, a reasonable strategy would be to deploy a tree-based ensemble as the main predictive model and use Elastic Net as a simpler companion model for explanatory analysis and robustness checks.

4.3.2 Model Results

The following section provides a detailed report of all the models estimated using the 2019–2022 data.

Let's denote:

- per_capita_DVMT = Per capita daily vehicle miles traveled
- pop_density = Population density (person/sq. mile)

- real_gdp_per_capita = Per capita Real GDP
- housing_units_per_1000 = Housing units per 1000 people
- pct_below_poverty = % population below poverty level (in fraction terms, e.g., 0.12 = 12%)
- agg_travel_time_per_employee = Aggregate travel time to work (minutes per employee)
- pct_commute_transit = % workers commuting by public transit (fraction 0–1)

OLS Coefficients and Statistical Significance

The estimated **OLS model** can be written as:

- per_capita_DVMT = -38.46 + 0.0005·pop_density + 0.0001· real_gdp_per_capita + 0.1550·housing_units_per_1000 + 9.4387·pct_below_poverty + 0.1156·agg_travel_time_per_employee - 247.8754·pct_commute_transit

Table 5 summarizes the OLS coefficient estimates, standard errors, t-statistics, p-values, and confidence intervals. Coefficients with p-values below 0.05 are conventionally considered statistically significant.

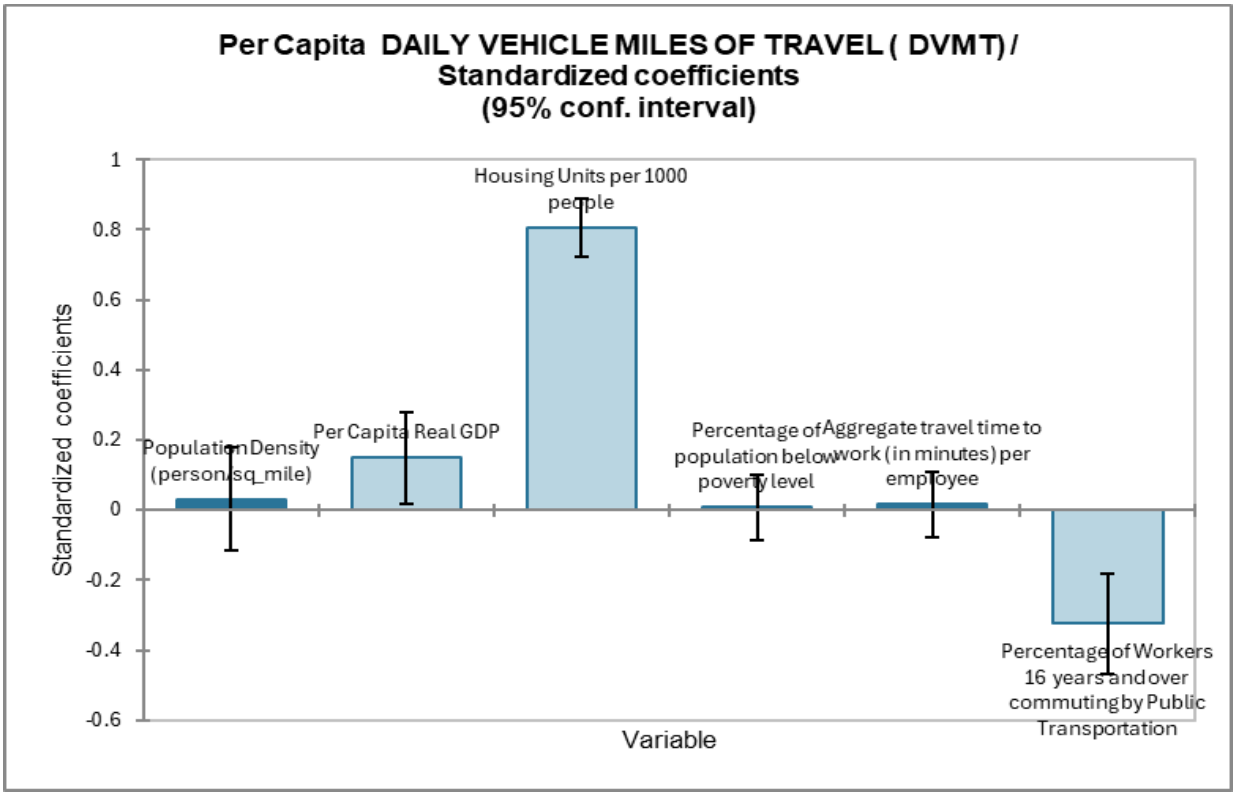
Table 5. The Estimated OLS Model

Variable	Coefficient	Standard Error	t	P>|t|	Lower Bound (95%)	Upper Bound (95%)	P-Values Significance Codes
Constant	-38.4646	14.5007	-2.6526	0.0086	-67.0391	-9.8901	**
Population Density (person/sq_mile)	0.0005	0.0011	0.4269	0.6699	-0.0016	0.0025	°
Per capita real GDP	0.0001	0.0001	2.2546	0.0251	0.0000	0.0002	.
Housing units per 1000 people	0.1550	0.0083	18.7788	0.0000	0.1388	0.1713	***
Percentage of population below poverty level	9.4387	58.0983	0.1625	0.8711	-105.0476	123.9251	°
Aggregate travel time to work (in minutes) per employee	0.1156	0.3501	0.3300	0.7417	-0.5744	0.8056	°
Percentage of workers 16 years and over commuting by public transportation	-247.8754	55.9595	-4.4295	0.0000	-358.1472	-137.6036	***

Signification codes: 0 < *** < 0.001 < ** < 0.01 < * < 0.05 < . < 0.1 < ° < 1

The standardized coefficients (Figure 14) provide insight into the relative importance of each variable.

Figure 14. Standardized Coefficients of the OLS Model



The OLS estimates indicate that three socioeconomic variables are significantly associated with county-level per capita DVMT after controlling for the full set of covariates. Housing units per 1000 people has the most pronounced positive influence on per capita DVMT. The raw coefficient (0.155) indicates that, holding all other factors constant, an increase of one housing unit per 1000 people is associated with an increase of approximately 0.155 miles per capita DVMT. This suggests that areas with a high ratio of housing stock to population—often characteristic of lower-density, sprawling development—experience higher rates of driving. Conversely, the percentage of workers commuting by public transportation yields a large, negative, and statistically significant coefficient (-247.8754). So, a 1 percentage-point increase in transit commute share is associated with a decrease of roughly 2.47 miles per capita DVMT, identifying public transit as a highly effective lever for reducing vehicle use. Per capita real GDP shows a small positive coefficient with borderline significance ($p = 0.025$); this finding is consistent with the decoupling hypothesis , suggesting a weakening link between economic output and driving. In other words, increases in economic activity no longer appear to translate into proportional growth in driving—a pattern widely observed across countries (Foster et al., 2021). The coefficients for population density, poverty rate, and commute time are not statistically significant at conventional levels.

Ridge, LASSO, and Elastic Net Model Estimation Results

Ridge, LASSO, and Elastic Net introduce regularization to stabilize coefficients in the presence of multicollinearity among predictors. Because all variables were standardized, coefficients can be interpreted as the expected change (in standard deviations) in per capita DVMT associated with a one-standard-deviation increase in each predictor, holding other variables constant.

The estimated model equations (standardized form) are:

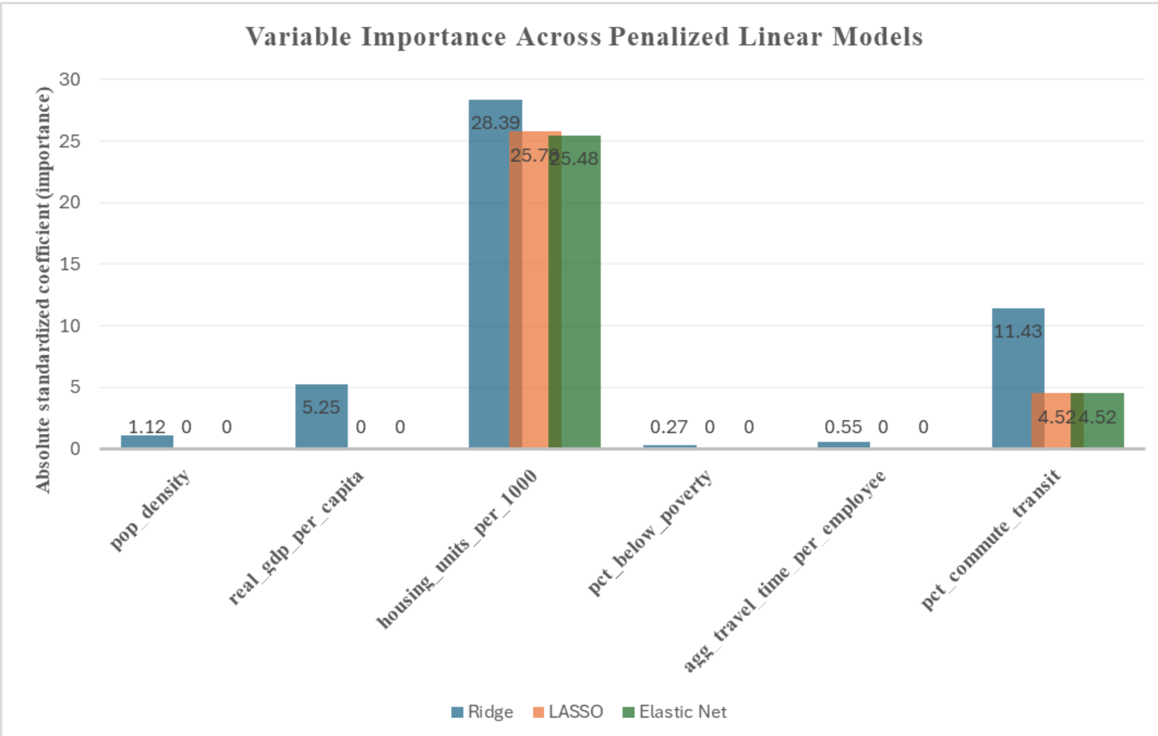
- **Ridge model:** per_capita_DVMT = 36.83+1.12·z(pop_density) + 5.25·z(real_gdp_per_capita) + 28.39·z(housing_units_per_1000) + 0.27·z(pct_below_poverty) + 0.55·z(agg_travel_time_per_employee) - 11.43·z(pct_commute_transit)

- **LASSO model:** per_capita_DVMT = 36.83+ 25.78·z(housing_units_per_1000) - 4.43·z(pct_commute_transit)

- **Elastic Net model:** per_capita_DVMT = 36.83 + 25.48·z(housing_units_per_1000) - 4.52·z(pct_commute_transit)

In these equations, per_capita_DVMT is in original units (miles) and z(·) denotes standardized variables. Ridge keeps all variables, but it shrinks some very close to zero. LASSO estimates coefficients with an L1 penalty and sets some coefficients exactly to zero, encouraging sparse solutions. The L1 penalty parameter (alpha) is selected via 5-fold cross-validation over a logarithmic grid (alpha = 2.09). Elastic Net combines L1 and L2 penalties. Hyperparameters are selected via 5-fold cross-validation (alpha = 1.91, l1_ratio = 0.99). It behaves similarly to LASSO but with slightly different magnitudes. Figure 15 shows the importance of each independent variable based on the standardized coefficients.

Figure 15. Variable Importance Across Ridge, LASSO, and Elastic Net Models



The penalized models broadly confirm the OLS findings: First, housing units per 1000 people exhibits a large, positive coefficient in all three models (approximately 25.48–28.39). This implies that counties with more housing stock per capita tend to have substantially higher per capita DVMT, even after controlling for other socioeconomic conditions. In standardized terms, a one standard deviation increase in this predictor is associated with an expected increase of approximately 25 additional miles of per capita DVMT, holding other predictors constant. This pattern is robust to the choice of penalty (Ridge, LASSO, Elastic Net), indicating that residential development intensity is the dominant predictor of vehicle travel in the models.

Second, the percentage of workers commuting by public transportation carries a negative coefficient in all models (roughly -4.52 to -11.43), and it is retained as a nonzero predictor in both LASSO and Elastic Net. This indicates that counties with higher transit commute shares tend to have lower per capita DVMT, consistent with mode substitution away from private vehicle travel. In the best-performing Elastic Net model, a one standard deviation increase in transit share corresponds to about an expected decrease of 4.52 miles in per capita DVMT.

By contrast, population density, per capita real GDP, poverty rate, and aggregate travel time to work have small coefficients in the Ridge model and are set exactly to zero in the LASSO and Elastic Net models. Once housing intensity and the share of transit commuters are included, these additional socioeconomic variables contribute little independent explanatory power to per capita

DVMT at the county level. Nevertheless, the signs of the coefficients generally align with the interpretation that higher per capita GDP, higher poverty level, and longer commute times are associated with higher vehicle travel demand. Collectively, the penalized regression results suggest that cross-county variation in per capita DVMT is driven primarily by the combination of housing stock and transit usage, with other socioeconomic characteristics playing a secondary role in this modeling framework.

Tree-Based Model Estimation Results

Tree-based models capture nonlinearities and interactions among the predictors. To interpret the trained models, feature importance measures are computed as the total reduction in squared error attributed to splits on each predictor, aggregated over all trees in the ensemble. The resulting importances are normalized to sum to 1. Figures 16 and 17 display the relative importance of each predictor for the Random Forest and Gradient Boosting models.

Figure 16. Random Forest Feature Importance

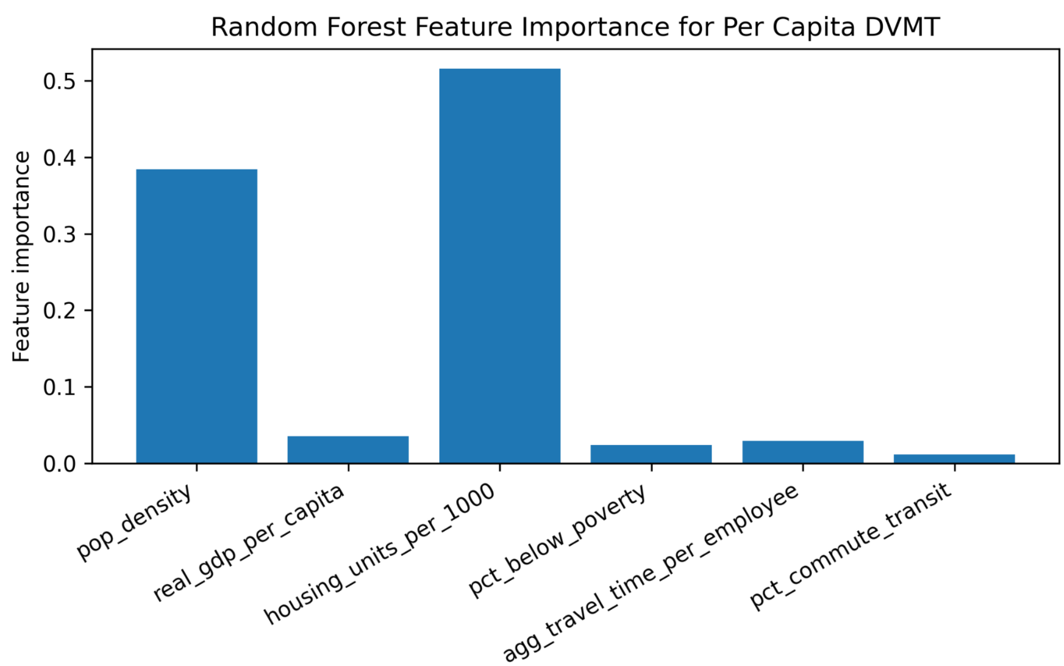
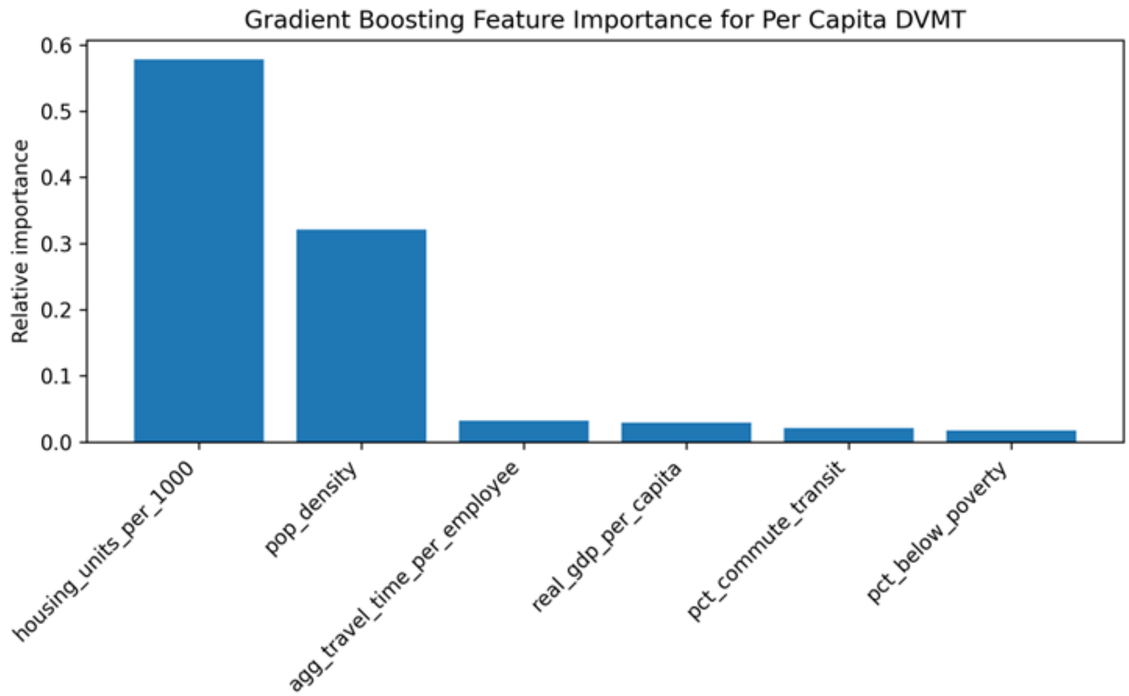


Figure 17. Gradient Boosting Feature Importance



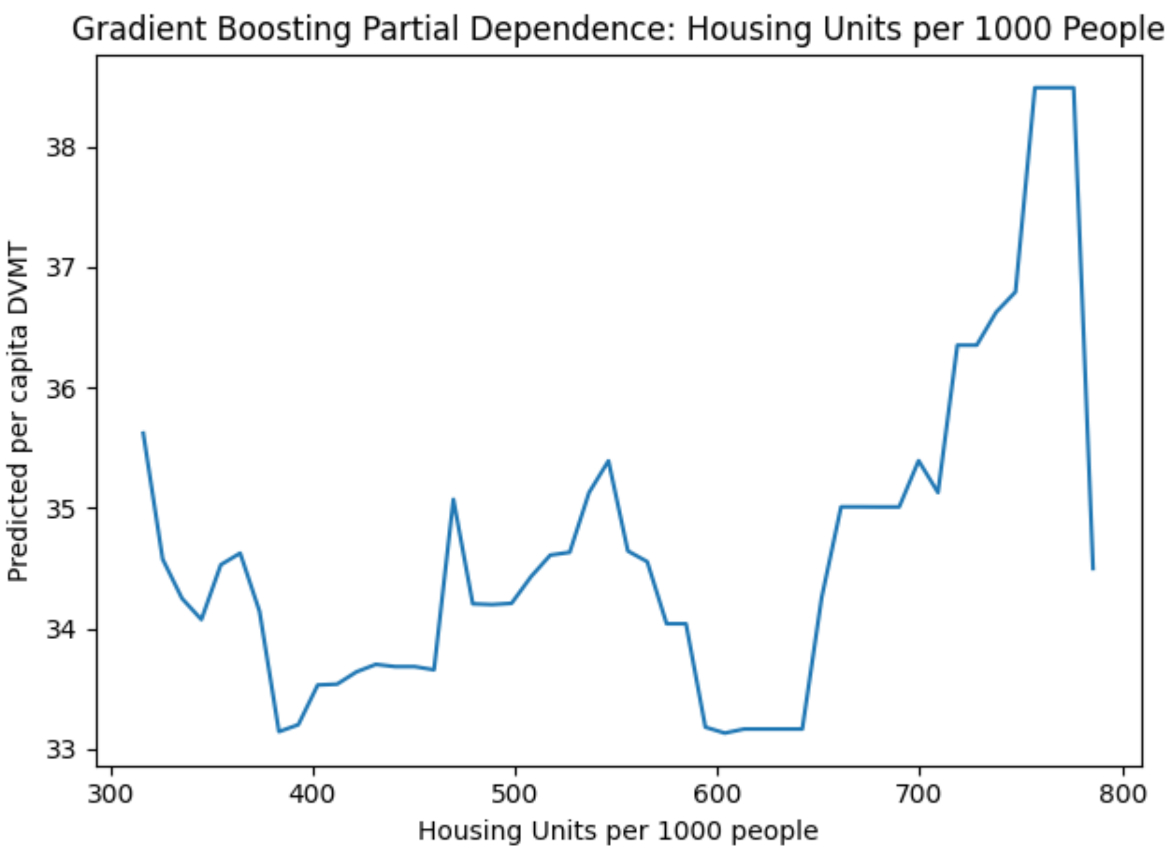
Across both tree-based models, the feature importance profiles tell a consistent story about which factors matter most for predicting per capita DVMT. In the Random Forest, housing units per 1000 people and population density dominate the ranking, together accounting for the vast majority of the overall importance, with housing intensity slightly ahead of population density. The remaining variables—real GDP per capita, aggregate travel time to work, poverty rate, and transit commute share—have much smaller, secondary contributions. This pattern suggests that the tree ensemble is relying primarily on indicators of the built environment and residential structure to partition the data and explain variation in per capita DVMT, using income and other socioeconomic measures mostly as fine-tuning variables rather than primary drivers.

The Gradient Boosting model reinforces and sharpens this hierarchy. It assigns an even larger share of importance to housing units per 1000 people, making it the unequivocal top predictor, with population density again emerging as the second most influential feature. The other four variables play only modest roles and, in the boosting model, their importances are even closer together and collectively much smaller than the top two. In practical terms, the comparison indicates that while both tree models see income, poverty, travel time, and transit share as relevant, they overwhelmingly attribute predictive power to how and where housing is supplied and how densely people are distributed. This convergence across two different nonlinear methods strengthens prior findings that land use and housing form are the central structural determinants of DVMT, with socioeconomic conditions and mode choice patterns acting mainly as modifiers around that core relationship (Zhang, et al., 2012; Nasri & Zhang, 2014; Ihlanfeldt, 2020; Tian et al., 2024).

Variables with larger importance scores are more influential in explaining cross-county variation in per capita DVMT. However, importance does not convey the direction of the association (positive or negative) or its functional form, which motivates the use of partial dependence plots (PDPs). PDPs show the marginal relationship between each predictor and the model's predicted per capita DVMT, averaging over the empirical distribution of all other variables. They help visualize whether associations are roughly linear, monotonic, or highly nonlinear.

To derive substantive, planning-relevant conclusions from the best-performing Gradient Boosting model, we analyze Partial Dependence Plots (PDPs) for the two dominant built environment variables: housing units per 1000 people and population density.

Figure 18. Partial Dependence Plot for Housing Units per 1000 People

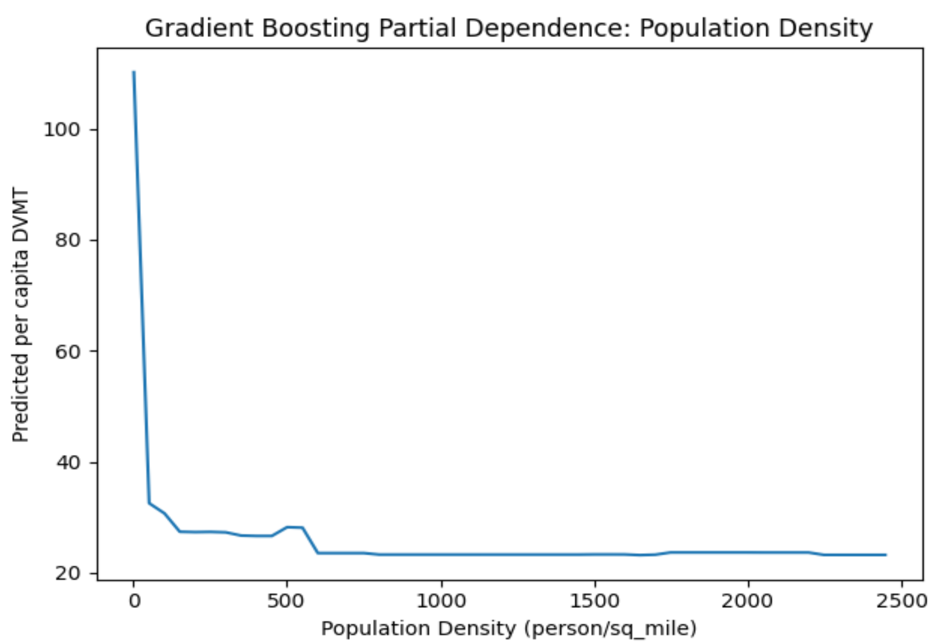


The PDP for housing units per 1000 residents (Figure 18) indicates a modest but substantively meaningful nonlinear relationship with predicted per capita DVMT. Across most of the observed range (approximately 350–650 units per 1000 residents), predicted per capita DVMT remains relatively stable in the low-to-mid 30s, with a shallow minimum of roughly 33–34 per capita DVMT at mid-range values of the housing-stock indicator. In this central band, variation in housing units per 1000 residents alone does not strongly alter predicted per capita DVMT once population density, income, poverty, commute time, and transit share are averaged over.

At the upper tail of the distribution, however, the PDP shows a clear increase: as housing units per 1000 residents move from roughly 650 to the upper end of the range (around 750–770 units), predicted per capita DVMT rises from the mid-30s to nearly 38.5 DVMT per capita. This pattern is consistent with the idea that counties characterized by a very high housing stock relative to their resident population—often reflecting suburban or second-home–intensive development—generate more vehicle travel per capita. These contexts likely combine smaller households, higher vehicle ownership per resident, and more dispersed residential patterns, all of which tend to increase daily driving. At the very lowest end of the housing-stock range, the PDP shows slightly elevated DVMT per capita that quickly declines into the mid-30s, suggesting that extremely low housing units per 1000 residents may also coincide with more rural conditions and longer trips, though this effect is modest relative to the high-end increase.

Overall, the magnitude of the PDP (about a 5-DVMT per capita swing from trough to peak) is smaller than the total DVMT per capita variation (7 to 287) in the data but is consistent with the high feature importance assigned to housing units per 1000 residents by the gradient boosting model. This indicates that the model uses this variable extensively, particularly to differentiate higher-DVMT per capita, housing-intensive counties from more moderate contexts.

Figure 19. Partial Dependence Plot for Population Density (Person/Sq_mile)



The PDP for population density (Figure 19) exhibits a much stronger nonlinearity and a wider vertical range in predicted DVMT. At extremely low densities (on the order of 2 persons per square mile), the model predicts very high per capita DVMT, on the order of 110 per capita DVMT, indicating extreme auto dependence in the sparsest rural counties. As density increases from these very low levels to approximately 100 persons per square mile, predicted DVMT drops sharply—from around 80–100 per capita DVMT down to roughly 30 per capita DVMT. This steep decline suggests that incremental increases in population density at the very low end substantially reduce the need for long-distance, high-frequency driving.

Beyond roughly 200–300 persons per square mile, the PDP flattens noticeably. Between about 600 and 2000 persons per square mile, predicted per capita DVMT hovers in a narrow band around 23–24 miles, and additional increases in density have only modest marginal effects. In other words, once a county reaches moderate urban or suburban densities, the gradient boosting model does not attribute large additional reductions in per capita DVMT to further densification alone. This plateau is consistent with the idea that, at the county scale, much of California's higher-density development remains car-oriented; further density may be necessary but not sufficient to substantially lower auto travel without concurrent changes in land use mix, transit supply, and other policies.

In the Gradient Boosting feature importance ranking, housing units per 1000 people was (slightly) more important than population density, even though the density PDP has larger vertical variation (from ~23 to 110), while the housing PDP spans only ~33 to 38.5. This is because population density's dramatic effect is concentrated in a tiny part of the space (the very lowest densities, represented by only a few rural counties). For most counties (say, population density > ~50–100), per capita DVMT is in a relatively narrow band and the population density PDP flattens out. Housing units per 1000 people, by contrast, is likely used in many trees across a wider portion of the sample to fine-tune splits between moderate and high DVMT counties, hence more overall contribution to error reduction even if its marginal PDP looks "smaller."

Taken together, the two PDPs underscore a coherent narrative. Population density primarily distinguishes extremely rural, auto-intensive counties from the rest of the distribution, with dramatic reductions in predicted per capita DVMT as density rises from very low to moderate levels. Within the non-extreme density range, however, housing units per 1000 people become more salient, with very high housing-stock intensity identifying counties where per capita DVMT is systematically higher even after conditioning on density and socioeconomic characteristics.

4.4 Policy Implications

Across all models, several consistent themes emerge:

A. Housing intensity and development pattern are central drivers of per capita DVMT. More housing units per 1000 people, typically reflecting more dispersed, auto-oriented development, are strongly associated with higher per capita DVMT.

B. Population density plays a complex but important role. While density is not statistically significant in the linear OLS specification, it is the most important predictor in both tree models. This suggests that density has a nonlinear relationship with per capita DVMT—very low densities may generate long-distance driving, whereas very high densities support shorter trips and alternative modes. Linear models struggle to capture this pattern.

C. Transit mode share is negatively associated with per capita DVMT. The OLS and penalized models indicate that higher shares of workers commuting by public transit are associated with lower per capita DVMT, conditional on other factors. Although the magnitude of this effect is reduced under regularization, its sign is stable.

D. Income and commute time play secondary roles. Per capita GDP has a small, positive, and only marginally significant effect in the linear model, while commute time and poverty rate show limited explanatory power after controlling for density, housing, and transit use.

From a policy standpoint, the results suggest that economic disincentives, such as tolls and fuel taxes, may be less effective in wealthier regions, where rising incomes enable drivers to absorb the additional costs. Consequently, strategies that alter land use patterns are likely to yield better results than socioeconomic shifts alone. Specifically, limiting low-density, auto-oriented development while increasing density in transit corridors appears to be a more robust approach. Furthermore, enhancing transit accessibility and mode share can reinforce these land use changes, thereby reducing vehicle use, particularly in denser contexts.

Overall, the results clearly demonstrate that nonlinear, tree-based machine learning models outperform traditional linear regression models in predicting per capita DVMT. These models suggest that per capita DVMT is best understood as a function of the built environment and transport system characteristics, with socioeconomic variables playing an important but secondary role. This is consistent with the empirical transportation literature and with emerging VMT-reduction policies that emphasize land use, density, and multimodal accessibility.

5. Summary and Conclusions

Vehicle miles traveled is a key transportation performance indicator that integrates demographic, economic, and behavioral factors influencing mobility, energy consumption, and infrastructure demand. This study presents an in-depth analysis and modeling of 2019–2023 DVMT and per capita DVMT at the county level in California. This research makes three contributions. First, through GIS mapping of DVMT changes and K-means clustering (k = 5), we developed a deep understanding of the spatial-temporal travel patterns and trends before, during, and after the COVID-19 pandemic. Second, using the Variance Inflation Factor (VIF) analysis to address multicollinearity, we tested a range of socioeconomic variables to identify the major predictors of vehicle travel. Third, we assessed the capability of linear and machine learning models to capture the spatial variability of per capita DVMT across California's diverse regions.

Six regression modeling approaches were evaluated: OLS, Ridge, LASSO, Elastic Net, Random Forest, and Gradient Boosting. The models were trained using six post-VIF explanatory variables: population density, per capita real GDP, housing units per 1000 people, percentage of the population below the poverty level, aggregate travel time to work per employee, and percentage of workers aged 16 years and over commuting by public transportation. The results clearly demonstrated that nonlinear, tree-based ensemble methods outperform traditional linear regression models in predicting per capita DVMT. Both the Random Forest and Gradient Boosting models explained over 98% of the variance in the training data. In contrast, the four linear regression models could only explain 59% to 66% of the variance. The analysis also revealed that housing units per 1000 people and population density were the dominant drivers of county-level per capita DVMT, whereas the contributions of public transit share, per capita real GDP, commuting time, and poverty level were secondary. These results confirmed that nonlinear relationships—particularly those involving interaction effects among built environment, population, vehicles, income, and economic activity—are fundamental to accurately modeling travel demand.

From a policy and planning perspective, these results have several implications.

First, time-series K-means clustering proved effective in uncovering archetypal mobility behaviors among California counties. High-growth clusters may require strategic investment in multimodal transportation, congestion management, and sustainability initiatives. Conversely, declining or stable clusters may indicate opportunities for adaptive reuse of roadway capacity or investment in green mobility options.

Second, the success of nonlinear ensemble methods highlights the potential of machine learning frameworks for regional transportation analysis and forecasting, particularly in contexts where traditional models oversimplify complex travel behavior.

Third, accurate county-level per capita DVMT predictions can inform infrastructure investment, emissions modeling, and land use planning, supporting California's broader transportation and climate goals under SB 375 and the Sustainable Communities and Climate Protection Act.

While this study provides a comprehensive spatial and modeling analysis of county-level DVMT and per capita DVMT in California, several limitations should be acknowledged. First, the analysis is conducted at the county level, which introduces aggregation bias. Counties vary substantially in geographic size, urban form, and internal socioeconomic heterogeneity. Relationships observed at the county scale may not reflect behavioral dynamics at finer spatial resolutions such as census tracts and neighborhoods. Second, although built environment and socioeconomic variables are strongly associated with per capita DVMT, the modeling framework establishes predictive rather than causal relationships. Omitted variables, such as fuel prices, electric vehicle adoption, roadway capacity, telework prevalence, etc., may also influence estimates. Third, while supported by silhouette metrics, the clustering results yield one cluster that encompasses the majority of counties. This may limit the segmentation granularity.

To enhance the evaluation and predictive power of this study, future research should extend the current framework by incorporating a broader set of exogenous variables. Specifically, incorporating fuel prices, evolving land use indicators, road capacity, and climate-related factors will enable more robust predictions of travel demand under dynamic demographic and policy scenarios. Furthermore, analysis of county-level DVMT changes (2019–2023) indicates that local factors, including tourism, wildfire recovery, network changes, EV adoption, and telework policies, significantly influence travel behavior. Quantitatively integrating these factors into the modeling would provide a more comprehensive assessment. Additionally, because current analyses often overlook local heterogeneity, there is a critical need to evaluate VMT patterns at finer geographic scales, such as the census tract or block group level. This granular approach will reveal localized disparities that regional averages often obscure. Ultimately, by bridging the gap between advanced predictive modeling and real-world policy application, California can move beyond reactive measures to proactively plan for a transportation system that ensures long-term sustainability, economic efficiency, and climate resilience.

Bibliography

Alexander, S. E., Alfonzo, M., & Lee, K. (2021). Safeguarding equity in off-site Vehicle Miles Traveled (VMT) mitigation in California. Mineta Transportation Institute. https://doi.org/10.31979/mti.2021.2027

Baek, S.-R. & Kim, D. (2023). A guide to navigating existing and emerging sources of local VMT and travel data. Mineta Transportation Institute. https://doi.org/10.31979/mti.2023.2232

Bamney, A., Jashami, H., Sonduru Pantangi, S., Ambabo, J., Megat-Johari, M.-U., Cai, Q., Gupta, N., & Savolainen, P. T. (2021). Examining impacts of COVID-19-related stay-at-home orders through a two-way random effects model. *Transportation Research Record: Journal of the Transportation Research Board*, *2677*(4), 255–266. https://doi.org/10.1177/03611981211046921

Blumenberg, E., & Pierce, G. (2012). Automobile ownership and travel by the poor: Evidence from the 2009 National Household Travel Survey. *Transportation Research Record: Journal of the Transportation Research Board*, *2320*(1), 28–36. https://doi.org/10.3141/2320-04 (Original work published 2012)

Blumenberg, E. & Siddiq, F. (2023). Commute distance and jobs-housing fit. *Transportation*, *50*, 869–891. https://doi.org/10.1007/s11116-022-10264-1

Boarnet, M. G., Bostic, R. W., Eisenlohr, A., Rodnyansky, S., Santiago-Bartolomei, R., & Jamme, H.-T. W. (2018). The joint effects of income, vehicle technology, and rail transit access on greenhouse gas emissions. *Transportation Research Record*, *2672*(24), 75–86. https://doi.org/10.1177/0361198118787087

Boarnet, M. G., & Wang, X. (2019). Urban spatial structure and the potential for vehicle miles traveled reduction: the effects of accessibility to jobs within and beyond employment sub-centers. *The Annals of Regional Science*. https://doi.org/10.1007/s00168-019-00900-7

Caltrans. (n.d.). Highway performance monitoring system (HPMS) data, retrieved October 14, 2024, from https://dot.ca.gov/programs/research-innovation-system-information/highway-performance-monitoring-system

Chi, B., & Lee, J. (2024). TOD effects on travel behavior: A synthesis of evidence from cross-sectional and longitudinal studies. *Journal of Transport and Land Use*.

Chung, S., Kwon, J., & Ahn, Y. (2024). Forecasting commercial Vehicle Miles Traveled (VMT) in urban California areas, Mineta Transportation Institute. https://doi.org/10.31979/mti.2024.2315

Circella et al. (2016) NCST white paper: What affects U.S. passenger travel? Current trends and future perspectives. Institute of Transportation Studies, University of California, Davis, Research Report UCD-ITS-RR-16-29.

Circella, G., Handy, S., & Boarnet, M. G. (2020). Impacts of employment density on passenger vehicle use and greenhouse gas emissions (Policy Brief). University of California, Davis, Institute of Transportation Studies.

Diao, M. i, & Ferreira, J. (2014). Vehicle Miles Traveled and the built environment: Evidence from vehicle safety inspection data. *Environment and Planning A: Economy and Space*, *46*(12), 2991–3009. https://doi.org/10.1068/a140039p

Dong, H. (2022). Can Californian households save money on transportation costs by living in Transit-Oriented Developments (TODs)? Mineta Transportation Institute. https://doi.org/10.31979/mti.2022.2012

Ewing, R., & Cervero, R. (2010). Travel and the built environment: A meta-analysis. *Journal of the American Planning Association*, *76*(3), 265–294. https://doi.org/10.1080/01944361003766766

Ewing, R., Hamidi, S., Gallivan, F., Nelson, A. C., & Grace, J. B. (2013). Combined effects of compact development, transportation investments, and road user pricing on Vehicle Miles Traveled in urbanized areas. *Transportation Research Record: Journal of the Transportation Research Board*, *2397*(1), 117–124. https://doi.org/10.3141/2397-14

Ewing, R., Hamidi, S., Gallivan, F., Nelson, A., Grace, C. (2014). Structural equation models of VMT growth in U.S. urbanized areas. *Urban Stud.*, *51*(14), 3079–3096.

Ewing, R., Hamidi, S., Grace, J. (2016). Compact development and VMT—environmental determinism, self-selection, or some of both? *Environ. Plan*, *43*(4), 737–755.

Federal Highway Administration. (2012). Exploring the relationship between travel demand and economic growth (FHWA-PL-12-026). Retrieved November 10, 2025, from: https://www.fhwa.dot.gov/policy/otps/pubs/vmt_gdp/index.cfm

Fisher, M. & LaMondia, J. (2021). Understanding the temporal, regional, demographic, and policy factors influencing counties' daily traffic volume reductions in response to

COVID-19. *Transportation Research Record: Journal of the Transportation Research Board*. 2677. 036119812110095. 10.1177/03611981211009541

Foster, V., Dim, J.U., Vollmer, S.J., & Zhang, F. (2021). Understanding drivers of decoupling of global transport CO2 emissions from economic growth: Evidence from 145 countries. *Policy Research Working Papers*.

Garceau, T. J., Atkinson-Palombo, C., & Garrick, N. (2014). Peak travel and the decoupling of vehicle travel from the economy: a synthesis of the literature. *Transportation Research Record*, *2412*(1), 41–48.

Gerayeli, A., & Jenkins, J. (2016). Land use and transportation modeling. *International Conference on Transportation and Development*. https://doi.org/10.1061/9780784479926.030

Goodwin, P., Dargay, J., & Hanly, M. (2004). Elasticities of road traffic and fuel consumption with respect to price and income: A review. *Transport Reviews*, *24*(3), 275–292. https://doi.org/10.1080/0144164042000181725

Governor's Office of Planning and Research (2018). Technical advisory on evaluating transportation impacts in CEQA. https://lci.ca.gov/docs/20180416-743_Technical_Advisory_4.16.18.pdf

Holtzclaw, J., Clear, R., Dittmar, H., Goldstein, D., & Haas, P. (2002). Location efficiency: Neighborhood and socio-economic characteristics determine auto ownership and use – Studies in Chicago, Los Angeles and San Francisco. *Transportation Planning and Technology*, *25*(1), 1–27. https://doi.org/10.1080/03081060290032033

Hong, J., Shen, Q., & Zhang, L. (2014). How do built environment factors affect travel behavior? A spatial analysis at different geographic scales. *Transportation*, *41*(3), 419–440.

Ihlanfeldt, K (2020). Vehicle miles traveled and the built environment: New evidence from panel data. *Journal of Transport and Land Use*, *13*(1). https://doi.org/10.5198/jtlu.2020.1647

Kalantari, H. A., Sabouri, S., Brewer, S., Ewing, R., & Tian, G. (2025). Machine learning in mode choice prediction as part of MPOs' regional travel demand models: Is it time for change? *Sustainability*, *17*, 3580. https://doi.org/10.3390/su17083580

Klakto, Trevor, Saeed, Tariq Usman, Volovski, Matthew, Labi, Samuel, Fricker, Jon, & Sinha, Kumares. (2017). Addressing the local-road VMT estimation problem using spatial interpolation techniques. *Journal of Transportation Engineering*, *143*. 10.1061/JTEPBS.0000064

Kim, S., Park, D., Heo, T.-Y., Kim, H., & Hong, D. (2016). Estimating vehicle miles traveled (VMT) in urban areas using regression kriging. *Journal of Advanced Transportation*, *50*(5), 769–785. https://doi.org/10.1002/atr.1374

Litman, T. (2025). Understanding transport demands and elasticities: How prices and other factors affect travel behavior. *Online Transportation Demand Management (TDM) Encyclopedia*, Victoria Transport Policy Institute, accessed November 12, 2025. http://www.vtpi.org/elasticities.pdf

Liu, F., Kaiser, R. G., Zekkos, M., & Allison, C. (2006). Growth forecasting of Vehicle Miles of Travel at county and statewide levels. *Transportation Research Record: Journal of the Transportation Research Board*, *1957*(1), 56–65. https://doi.org/10.1177/0361198106195700109

McMullen, B. S., & Eckstein, N. (2013). Determinants of VMT in urban areas: A panel study of 87 U.S. urban areas 1982–2009. *Journal of the Transportation Research Forum*, *52*(3), 5–24. https://doi.org/10.5399/osu/jtrf.52.3.4177

Millard-Ball, A., & Schipper, L. (2010). Are we reaching peak travel? Trends in passenger transport in eight industrialized countries. *Transport Reviews*, *31*. https://doi.org/10.1080/01441647.2010.518291

Millard-Ball, A., & Rosen, M. (2025). *Road capacity as a fundamental determinant of vehicle travel*. https://doi.org/10.7922/G2Q52N05

Mondschein, A., & Taylor, B. (2017). *Low-income households and VMT: An equity analysis of transportation costs and travel behavior*. University of California Institute of Transportation Studies.

Nasri, A., & Zhang, L. (2014). Assessing the impact of metropolitan-level, county-level, and local-level built environment on travel behavior: Evidence from 19 U.S. urban areas. *Journal of Urban Planning and Development*, *141*, 04014031. 10.1061/(ASCE)UP.1943-5444.0000226

Pickrell, D., Pace, D. M., Wishart, J., & Roycroft, C. (2023). Development of VMT forecasting models for use by the Federal Highway Administration, Office of Highway Policy Information. Federal Highway Administration. Retrieved November 14, 2025, from https://www.fhwa.dot.gov/policyinformation/tables/vmt/vmt_model_dev.cfm

Reiffer, A., Magdolen, M., Ecke, L., & Vortisch, P. (2023). Effects of COVID-19 on telework and commuting behavior: Evidence from 3 years of panel data. *Transportation Research Record*, *2677*(4), 478–493. https://doi.org/10.1177/03611981221089938

Salon, D., Boarnet, M. G., Handy, S., Spears, S., & Tal, G. (2012). How do local actions affect VMT? A critical review of the empirical evidence. *Transportation research part D: transport and environment*, *17*(7), 495–508.

Salon, D. (2016). *Socioeconomic status, vehicle ownership, and travel behavior in California*. Mineta Transportation Institute.

Sabouri, S., Ewing, R., & Kalantari, H.A. (2024). Estimating transit's land use multiplier: direct and indirect effects on vehicle miles traveled. *Transportation*. https://doi.org/10.1007/s11116-024-10542-0

Sardari, R., Hamidi, S., & Pouladi, R. (2018). Effects of traffic congestion on Vehicle Miles Traveled. *Transportation Research Record: Journal of the Transportation Research Board*, *2672*(47), 92–102. https://doi.org/10.1177/0361198118791865

Schaber, K., Arambepola, R., Schluth, C., Labrique, A. B., Mehta, S. H., Solomon, S. S., Cummings, D. A. T., & Wesolowski, A. (2024). Geography versus sociodemographics as predictors of changes in daily mobility across the USA during the COVID-19 pandemic: A two-stage regression analysis across 26 metropolitan areas. *BMJ Open*, *14*(7), e077153. https://doi.org/10.1136/bmjopen-2023-077153

Su, Rongxiang, Xiao, Jingyi, Shi, Hui, & Goulias, Konstadinos. (2022). Heterogeneity in the Relationship between Vmt and the Built Environment: A quantile regression approach. *SSRN Electronic Journal*. 10.2139/ssrn.4144179.

Tapio, P. (2005). Towards a theory of decoupling: Degrees of decoupling in the EU and the case of road traffic in Finland between 1970 and 2001. *Transport Policy*, *12*, 137–151. https://doi.org/10.1016/j.tranpol.2005.01.001

Tian, G., Danton, B., Li, B., Gopu, V., & Codjoe, J. A. (2024). Understanding household VMT generation: A comparative analysis with traditional statistical models and a machine-learning approach. *The Journal of Transport and Land Use*, *17*(1): 881–901.

Urban Displacement Project. (2021). Assessing the travel demand and co-benefit impacts of affordable housing near transit (Final report). Retrieved November 10, 2025, from https://www.urbandisplacement.org/research/assessing-the-travel-demand-and-co-benefit-impacts-of-affordable-transit-oriented-developments/

Williams, T. A., Chigoy, B., Borowlec, J., & Glover, B. (2016). Methodologies used to estimate and forecast vehicle miles traveled (VMT): Final report. Texas A&M Transportation Institute, PRC 15-40 F. https://rosap.ntl.bts.gov

Woldeamanuel, M., & Kent, A. (2014). Determinants of per capita Vehicle Miles Traveled (VMT): The case of California. *Journal of the Transportation Research Forum*, *53* (3), 35–46.

Zhang, L., Hong, J., Nasri, A., & Shen, Q. (2012). How built environment affects travel behavior: A comparative analysis of the connections between land use and Vehicle Miles Traveled in U.S. Cities. *Journal of Transport and Land Use*, *5*, 40–52. 10.5198/jtlu.v5i3.266

Zheng, Y., Wang, S., Liu, L. et al. (2024). Impacts of remote work on vehicle miles traveled and transit ridership in the USA. *Nat Cities*, *1*, 346–358. https://doi.org/10.1038/s44284-024-00057-1

Zhou, Y., Aeschliman, S., & Gohlke, D. (2021). Household transportation energy affordability by region and socioeconomic factors. *Transportation Research Record: Journal of the Transportation Research Board*, *2675*(10), 81–95. https://doi.org/10.1177/03611981211010186

About the Authors

Yong Lao, PhD

Yong Lao is a Full Professor of Geography in the Department of Social Sciences and Global Studies at California State University (CSU), Monterey Bay. Dr. Lao's teaching and research interests are Geographic Information Systems (GIS), transportation planning, location theories, quantitative methods, and economic geography. He has collaborated with many government agencies and businesses on a wide range of GIS mapping and spatial analysis projects, including performance evaluation of bus lines, hazardous-material shipping, childcare facility selection, airline hub location, and ATM network planning.

Bo Yang, PhD

Bo Yang is an Assistant Professor of Environmental Studies at the University of California, Santa Cruz. He is the Director of the GISTAR M.A. program and leads the GeoFly Lab, where his research focuses on GIS, remote sensing, UAV/drone mapping, and geospatial AI for environmental and societal applications. His work spans wildfire risk assessment, transportation and infrastructure monitoring, and coastal ecosystem mapping, and has been supported by agencies including the NSF, NASA, USDOT, and CAL FIRE.

MTI FOUNDER

Hon. Norman Y. Mineta

MTI BOARD OF TRUSTEES

Directors